

Citations, Citation Indicators, and Research Quality: An Overview of Basic Concepts and Theories

SAGE Open
January-March 2019: 1–17
© The Author(s) 2019
DOI: 10.1177/2158244019829575
journals.sagepub.com/home/sgo
 SAGE

Dag W. Aksnes¹ , Liv Langfeldt¹, and Paul Wouters²

Abstract

Citations are increasingly used as performance indicators in research policy and within the research system. Usually, citations are assumed to reflect the impact of the research or its quality. What is the justification for these assumptions and how do citations relate to research quality? These and similar issues have been addressed through several decades of scientometric research. This article provides an overview of some of the main issues at stake, including theories of citation and the interpretation and validity of citations as performance measures. Research quality is a multidimensional concept, where plausibility/soundness, originality, scientific value, and societal value commonly are perceived as key characteristics. The article investigates how citations may relate to these various research quality dimensions. It is argued that citations reflect aspects related to scientific impact and relevance, although with important limitations. On the contrary, there is no evidence that citations reflect other key dimensions of research quality. Hence, an increased use of citation indicators in research evaluation and funding may imply less attention to these other research quality dimensions, such as solidity/plausibility, originality, and societal value.

Keywords

citations, indicators, metrics, bibliometrics, evaluation, research quality

Introduction

In recent years, bibliometric indicators have increasingly been applied in the context of research evaluation as well as research policy more generally. Examples include the use of citation indicators in evaluation of the scientific performance of research groups, departments, and institutions (Moed, 2005); evaluation of research proposals (Cabezas-Clavijo, Robinson-Garcia, Escabias, & Jimenez-Contreras, 2013); allocation of research funding (Carlsson, 2009); and hiring of academic personnel (Holden, Rosenberg, & Barker, 2005). Citation measures are also core indicators in several university rankings, such as the Leiden ranking and Academic Ranking of World Universities (ARWU) (Piro & Sivertsen, 2016).

Thus, indicators or metrics are applied for a variety of purposes and have permeated many aspects of the research system. Traditionally, peer review has been the “gold standard” for research assessment. Increasingly, metrics are being applied as an alternative, by its own or in combination with peer review. For example, citation data were used in the United Kingdom to inform their peer-review judgments by some panels in the 2014 Research Excellence Framework (REF; Wilsdon et al., 2015). This raises the question of the

reliability and validity of citations as performance indicators. In which contexts and for which purposes are they suitable? These are questions which have been debated over the past decades.

In the most radical version, it has been argued that assessment of research based on citations and other bibliometric measures is superior compared with the traditional peer-review method. For example, Abramo and D’Angelo (2011) claimed,

Empirical evidence shows that for the natural and formal sciences, the bibliometric methodology is by far preferable to peer-review. . . . Compromise methods, such as informed peer review, in which the reviewer can also draw on bibliometric indicators in forming a judgment, do not, in the opinion of the authors, offer advantages that justify the additional costs:

¹Nordic Institute for Studies in Innovation, Research and Education (NIFU), Oslo, Norway

²Centre for Science and Technology Studies (CWTS), Leiden University, The Netherlands

Corresponding Author:

Dag W. Aksnes, Nordic Institute for Studies in Innovation, Research and Education (NIFU), P.O. Box 2815, Tøyen, Oslo 0608, Norway.
Email: Dag.W.Aksnes@nifu.no



indicators will not assist in composing human judgments, at the maximum permitting a confirmation or refutation. (p. 512)

Similar viewpoints have been put forward by Regibeau and Rockett (2016).¹

Nevertheless, the application of bibliometric indicators for assessing scientific performance has always been controversial. For a long time, the use of journal impact factors (JIFs) in research evaluation contexts has been heavily criticized (Cagan, 2013; Hicks, Wouters, Waltman, de Rijcke, & Rafols, 2015; Seglen, 1989). Moreover, the application of citation indicators has also been criticized more generally, with respect to their validity as performance measures and their potentially negative impact upon the research system (MacRoberts & MacRoberts, 1989; Osterloh & Frey, 2015; Weingart, 2004). For example, Seglen (1998) examined problems attached to citation analyses and concluded that "... citation rates are determined by so many technical factors that it is doubtful whether pure scientific quality has any detectible effect at all..." (p. 226)

Broadly speaking, while extensive discussions appeared during the 1970s and 1980s on what citations actually "measure" and how citations relate to scientific quality (see, for example, Cronin, 1984), this issue seems to have received less attention in recent decades. Nowadays, it is often taken for granted that citations in some way measure scientific impact, one of the constituents of the concept of scientific quality. More attention has been paid to methodological issues such as appropriate methods for normalizing absolute citation counts (Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011b), in addition to development and examinations of new citation-based indicators such as the h-index (Bornmann & Daniel, 2007; Waltman, 2016). Although the latter development has contributed to important progress in the field, the limitations of citations discussed in the 1970s and 1980s did not disappear. In the scientific paper, the references have various purposes. Authors are not including references merely because of their scientific quality. The selection of references is determined by various factors, one being their relevance for the research topic being addressed (Bornmann & Daniel, 2008). These limitations cannot be overcome by the construction of technically more sophisticated or reliable indicators.

Against this background, this article provides an overview of basic issues related to citations, citation indicators, and their interpretation and validity as performance measures.² The question of how citations may relate to or reflect various aspects of the concept of research quality is paid particular attention. The research literature on these topics is huge, covering numerous issues and research questions. This article is written as an introductory overview for a broader audience interested in these topics. Therefore, the coverage of topics and literature is selective and does not discuss all details. In addition, the literature on the interaction between citing practices and evaluation processes is only referred to

in passing, and we do not discuss constructivist and semiotic theories of quality and citation (Wouters, in press).

The article is structured as follows: As an introduction, we describe some basic issues relating to the construction of citation indicators. The "Citation Indicators" part focuses on the citation process and which roles the references have in the scientific paper. Many previous studies have compared citation indicators with the outcome of peer review, and in the "Understanding Citations" part, this issue is examined. Some factors affecting the validity of citation indicators are further described in "Validation Studies" part. In the "Citations as Indicators—Other Validity Issues" part, the question concerning citations and the concept of research quality is addressed. Research quality is a multidimensional concept. Therefore, we discuss how citations may relate to each of the various dimensions of the quality concept. While the first to the fourth part provide a condensed review of the issues at stake, the last part is more explorative and discursive. The reason is that few previous studies have addressed the topic systematically.

Citation Indicators

The development of bibliometrics as a field is strongly linked to the creation of the *Science Citation Index* (SCI) by Eugene Garfield in 1961 (Aksnes, 2005). Originally, this bibliographic database was mainly constructed for information retrieval purposes, to aid researchers in identifying relevant articles in the huge research literature archives (Welljams-Dorof, 1997). As a supplemental property, it enabled scientific literature to be analyzed quantitatively. Since the 1960s, the SCI and other similar databases, now included in the online product *Web of Science*, have been applied in a large number of studies covering many different fields. The option for citation analysis has been a crucial cause for this popularity (Aksnes, 2005). In the database, all the references of the indexed articles are registered. Based on this, each article can be ascribed a citation count showing how many times it has been cited by later papers registered in the database. Citation counts and indicators can then be calculated for aggregated "publication levels," for example, representing research units, departments, or scientific fields. In the early 2000s, competing databases were introduced which also include citation statistics, most importantly the *Scopus database* (launched in 2004) and *Google Scholar* (launched in 2004). The coverage of the scientific and scholarly literature varies across these databases, and the results of citation studies are thus dependent upon the particular characteristics of the databases and their coverage.

During recent decades, a large number of different citation indicators have been developed and there has been extensive debate about appropriate methods for calculating citation indicators, normalization procedures, database coverage, and data quality (for an overview, see de Rijcke, Wouters, Rushforth, Franssen, & Hammarfelt, 2016; Moed,

2005; Vinkler, 2010; Waltman, 2016). Among the most frequently used citation indicators are the field-normalized citation impact indicator, the number/proportion of highly cited papers, and the h-index. The first indicator is an expression of the average number of citations of the publications, normalized for field, publication year, and document type (e.g., regular article or review). For example, a value of two tells us that the publications have been cited twice above the average of their field and publication year, that is, the world average (Waltman et al., 2011b). Indicators relating to highly cited papers are typically percentile-based, for example, the number and proportion of publications that belong to the top 1% or top 10% most frequently cited of their fields (adjusted for publication year; Waltman & Schreiber, 2013). Another citation-based indicator is the JIF which, despite problems, flaws, and recommendation for not using it in research evaluation contexts, continues to be a very popular bibliometric indicator if not the most popular one (Bornmann, Marx, Gasparyan, & Kitas, 2012; Cagan, 2013).

There are large variations in average citation rates across different subject areas. For example, in many humanities disciplines, an average paper receives less than one citation during a 10-year period, compared with more than 40 citations in some biomedical fields (data from Web of Science 2005-2015). According to Marx and Bornmann (2015), the main reason for such differences relates to the coverage of the database. Only a small fraction of the scholarly literature in the humanities is represented in the Web of Science, and most of the references and citations will not be captured by the database. Accordingly, the average citation rate within the humanities is much higher when using other databases which cover the literature better, such as Google Scholar (Harzing & Alakangas, 2016). In addition, the average number and age of the references, and the ratio of new publications in the field and the total number of publications play a role when it comes to field differences in citation rates (Aksnes, 2005).

Because there are large field and temporal differences in how many citations an average paper receives, it was suggested in the early days of scientometrics that the absolute citation counts need to be normalized (Schubert & Braun, 1986; Schubert, Glänzel, & Braun, 1987).³ It has since been the standard to adjust for field, publication year, and publication type when calculating citation indicators. The most commonly known indicator is the field-normalized citation impact indicator, previously known as the crown indicator (van Raan, 2004) where the above-mentioned differences are taken into account. By this indicator, one attempts to correct for the effect of the variables, which are considered to be disturbing factors in citation analyses (i.e., associated with imbalance in citation opportunities). In recent years, much attention has been devoted to methods for normalization, to the question of how to delineate scientific fields used in the normalization and whether the normalizations should be carried out at the level of individual paper or at aggregated

paper levels (averages of ratios [AoR] vs. ratios of averages [RoA]; Opthof & Leydesdorff, 2010; Waltman & van Eck, 2013). There is no general agreement on what is the most appropriate method (Ioannidis, Boyack, & Wouters, 2016), but empirical studies have shown that two different methods for normalization, AoR and RoA, did not produce very different results, particular at the level of countries and institutions (Larivière & Gingras, 2011; Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011a).

Citation distributions are very skewed. This skewness was already identified by the historian of science Derek de Solla Price (1965). The larger part of all scientific papers are never cited or cited a few times only in the subsequent scientific literature (Aksnes, 2005). On the contrary, some articles have an extremely large number of citations, reaching into the hundreds and even thousands. During the recent two decades, there has been growing interest toward using the top end, the highly cited papers, as performance indicators. The expectation is that these papers represent extra-ordinarily good work and hence may be used to identify scientific excellence, an increasing concern in science policy (Langfeldt et al., 2015; van Raan, 2000). There are different types of such indicators; a common indicator is the number or proportion of articles that belong to the top 1% or 10% most frequently cited papers (in the same field and in the same year).

The h-index was introduced in 2005 (Hirsch, 2005) and rapidly became a very popular bibliometric measure. This indicator takes both the number of articles produced and the citation impact of these articles into account. According to the definition of the h-index, a researcher with an h-index of 15 has at least 15 publications with at least 15 citations. The index was originally developed for analysis of individuals, but has also been applied at other levels, such as research groups, departments, and institutions. Despite its popularity, the indicator has several problems. Most importantly, it is not field-normalized and no corrections are made for career length, which means that the indicator disfavors younger researchers (for a review, see, for example, Alonso, Cabrerizo, Herrera-Viedma, & Herrera, 2009).

When measuring citation frequencies, the temporal dimension or time window is important. Usually, articles that have been published recently have hardly been cited yet and the number of citations increases over time as older papers have had more time to accrue citations. In citation analyses, various time windows are used, depending on the purpose and field analyzed. Frequently, a citation window of 3 to 5 years is used (Council of Canadian Academies, 2012). This is a pragmatic compromise between a short- and long-term citation window (Leydesdorff, Wouters, & Bornmann, 2016). However, the extent to which short-term citation rates can be considered as predictor of long-term rates will vary (Baumgartner & Leydesdorff, 2014) and using short-term windows (e.g., 2 or 3 years) means that contributions to the current research front are appreciated more than long-term impact (Leydesdorff et al., 2016). A longer citation window is usually considered as

more reliable than shorter windows. For example, Levitt and Thelwall (2011) have argued that short citation windows have the problem that articles published earlier in a year have a significant advantage (i.e., are on average more highly cited) compared with publications appearing later in a year. On the contrary, a disproportionate long-time period makes the results less usable for evaluation purposes. The reason is that one then only has citation data available for articles published many years previously (Aksnes, 2005). For instance, applying a citation window of 3 years means that articles need to be at least 3 years old to be included in the analysis. Thus, contributions from the most recent years, the period which would typically be of particular interest in research assessment exercises (RAEs), cannot be assessed.

Understanding Citations

The question of what citations “measure” has for a long time been an important question in bibliometrics. Two of the pioneers within citation studies, the Cole brothers, often referred to citations as a measure of quality, although a slightly more cautious definition was given in the introduction of their book on social stratification in science: “The number of citations is taken to represent the relative scientific significance or ‘quality’ of papers” (J. R. Cole & Cole, 1973, p. 21). Even today, citation indicators are sometimes presented as measures of scientific quality (see, for example, Abramo & D’Angelo, 2011; Durieux & Geveno, 2010).

Because citations are derived from the references in the literature, it has been a common assumption that the use of citations as research performance indicators should be justified or grounded in the referencing behavior of authors. Already in 1981, Smith complained,

Not enough is known about the “citation behavior” of authors—why the author makes citations, why he makes his particular citations, and how they reflect or do not reflect his actual research and use of the literature. When more is learned about the actual norms and practices involved, we will be in a better position to know whether (and in what ways) it makes sense to use citation analysis in various application areas. (p. 99)

Many studies on referencing behavior have indeed been conducted. We refer to Bornmann and Daniel (2008) and Nicolaisen (2007) for extensive overviews of this literature. More recent contributions include, for example, Camacho-Minano and Nunez-Nickel (2009), Thornley et al. (2015), and Willett (2013). Roughly speaking, two contrasting perspectives may be identified: one in which the intellectual function of the references is emphasized and one analyzing citing as fundamentally a social process. Typically, the latter approach would focus on “outside” and social factors rather than content, and has mostly been associated with attempts to critique the use of citations as performance measures (Aksnes, 2005).

The Role of References in the Scientific Paper

Studies undertaken have revealed that the role of the reference, both in the citing text and with respect to the cited text, is complex. For example, already in 1964, Garfield suggested 15 different reasons for why authors cite other publications (reprinted in Garfield, 1977). Among these were providing background reading, identifying methodology, paying homage to pioneers, identifying original publication or other work describing an eponymic concept, identifying original publications in which an idea or concept was discussed, giving credit for related work, criticizing previous work, correcting a work, substantiating claims, alerts to a forthcoming work, providing leads to poorly disseminated work, authenticating data and classes of fact—physical constants and so on—disclaiming works of others, and disputing priority claims.

Hence, the textual functions of citations vary considerably. In a scientific article, some of the references will represent works that are crucial or significant antecedents to the present work; others may represent more general background literature (Aksnes, 2005). For example, in a review of the literature published on the topic during 1965–1980, Small (1982) identified five distinctions: A cited work may be (a) refuted, (b) noted only, (c) reviewed, (d) applied, or (e) supported by the citing work. These categories were respectively characterized as (a) negative, (b) perfunctory, (c) compared, (d) used, and (e) substantiated. This means that the different functions the references may have in a text are much more complex than merely providing documentation and support for particular claims.

These and later studies have revealed that the references have a multitude of functions in the scientific article. With respect to the relation between citation frequency and scientific quality, patterns at aggregated levels are relevant to consider, not only the individual articles. To explain how some papers come to be highly cited, one has to focus on how references at micro-levels aggregate (Aksnes, 2005). Typically, a scientific article is structured as a progression from the general to the particular (Law, 1986). This means that the introduction of an article typically contains references to more general or basic works within a field. The accumulative effect of many articles referring to the same general works is that such contributions get a very large number of citations. References to highly cited publications are more often present in the introduction than in other parts of the publications (Voos & Dagaev, 1976).

Correspondingly, most scientific publications contain a methods section in which the methods applied in the study are documented. Here, authors typically cite the basic papers describing these methods. Because of this, articles describing commonly used methods may receive a very large number of citations. The prime example here is an article from 1951 on protein measurement (Lowry, Rosebrough, Farr, & Randal, 1951), which is the most highly cited paper ever.

This article has now been cited more than 305,000 times in the Web of Science database (Van Noorden, Maher, & Nuzzo, 2014).

Although important insights on the role of references in the scientific article have been obtained, the accumulation of knowledge at the same time has been hampered by the fact that different classification systems have been applied in previous studies (Liu, 1993). Moreover, the studies are often based on rather small samples of papers from selected scientific fields, and the results may not have general validity. According to Bornmann and Daniel (2008), many studies have methodological weaknesses and have provided findings with little reliability.

Citation Behavior

Robert K. Merton is often considered to have provided the original theoretical basis for linking citations counts to the use and quality of the scientific contributions (Aksnes, 2005). According to Merton's view, the norms of science oblige researchers to cite the work upon which they draw, and in this way acknowledge or credit contributions by others (Merton, 1979). Such norms are maintained through informal interaction in scientific communities and through peer review of submitted manuscripts. If authors cite the works they find useful, frequently cited publications may be assumed to have been more useful than papers which are hardly cited at all. Thus, the number of citations may be regarded as a measure of the usefulness, impact, or influence of a publication. The same reasoning can be used for aggregated levels of publications. The more citations the publications of, for example, a department draw, the greater their influence must be. There are also discipline-specific norms or even codes that differ by journal within a field, for example, concerning how and when to cite, and how many references a paper should contain (Hellqvist, 2010).

Empirical studies have shown that the Mertonian account of the normative structure of science covers only part of the dynamics (Aksnes, 2005). For the citation process, this implies that other incentives shape the citing patterns, like creating visibility for one's work through self-citations or citing a journal editor's work as an attempt to enhance the chances of acceptance for publication. Previous studies have revealed a multitude of motivations, functions, and causes of references in scientific communication (Bornmann & Daniel, 2008).

Early contributions addressing the social dimensions of the references were made by Gilbert and later by MacRoberts and MacRoberts and others. Gilbert (1977) argued that citing ("referencing") is essentially a device for persuasion. To persuade the scientific community of the value and importance of their publication, authors are using references as rhetorical tools. References vary in their power of persuasion. Therefore, it will be more persuasive to cite an authoritative

paper, and authors tend to select references that will be regarded as authoritative by the intended audience.

Moreover, characteristics of authors' referencing behavior have been used for arguing against the use of citations as performance indicators, for example, by MacRoberts and MacRoberts (1989, 1996). Based on empirical case studies, they showed that a very small proportion of the knowledge basis of an article (consisting of hundreds or thousands of former publications) actually are cited. Moreover, the citing is biased: some sources are cited essentially every time they are used, while other research is never cited even though it may be used more often than the highly cited work. Accordingly, they criticize citation analysts who

in spite of an overwhelming body of evidence to the contrary . . . continue to accept the traditional view of science as a privileged enterprise free of cultural bias and self-interest and accordingly continue to treat citations as if they were culture free measures. (MacRoberts & MacRoberts, 1996, p. 442)

The views of the MacRoberts's previously led to much debate, but their conclusions are generally seen as too sweeping (Aksnes, 2005). Garfield, for example, claimed that it would be impossible to cite all former literature on a particular topic. According to the founder of the SCI, the fact that authors do not cite all their influences does not invalidate the use of citations as performance measures when enough literature is taken into account (see Garfield, 1997). Although most citation analysts seem to agree that citing or referencing is biased, it has been argued that this bias is not fatal for the use of citation as performance indicators—to a certain extent, the biases are averaged out at aggregated levels. According to Luukkonen (1990), the presence of different cognitive meanings of citations and motivations for citing does not necessarily invalidate the use of citations as (imperfect) performance measures. Motives and consequences are analytically distinct.

Still, the different approaches need not preclude each other. In fact, some authors have tried to develop a multidimensional approach (Amsterdamska & Leydesdorff, 1989; Cozzens, 1989; Glaser & Laudel, 2001; Leydesdorff, 1989; Luukkonen, 1997b). Cozzens, for example, has emphasized that a pluralistic explanation of citations means that we accept aspects of all perspectives. In the course of writing a paper, a scientist's actions may be oriented to one or another aspect. On one hand, the citation behavior of individuals is affected by external pressures and there are personal motives, self-interests, and so forth in the citation process; on the other, there are certain norms, rules, traditions, and etiquettes that limit the scope and acceptability of individual actions. Thus, there are rules for behavior and interaction, even if not the traditional Mertonian ones. Instead of standard ("ideal") versus deviation, an interesting question is to understand the patterns, and perhaps identify ways to link quality to particular features of citation processes.

Aksnes (2003) introduced a conceptual distinction between quality dynamics and visibility dynamics to explain how micro-level decisions to cite particular papers aggregate and result in highly cited publications. Here, the quality dynamic is grounded in the structure of scientific knowledge. Typically, scientific progress is achieved through a variety of contributions. Some represent major scientific advances; others are filling in the details. This distinction is related to Cole's concepts of core and frontier knowledge (S. Cole, 1992). In the view of Cole, core knowledge consists of the basic theories within a field, while frontier knowledge is knowledge currently being produced. Much of the research produced at the frontier are low-level descriptive analyses or represent contributions that turn out to be of little or no lasting significance (S. Cole, 2000). Therefore, a large part of what is published does not as such pass its way into core knowledge. Also, parts of what is published represent "dead ends" and does not function as a basis for further knowledge development. In consequence, according to Aksnes (2003), one expects a skewed distribution of citation scores and differences between fields depending on the relationship between evolving core knowledge and more ephemeral frontier knowledge. At the same time, citation frequencies are determined by other mechanisms and are not a simple reflection of the quality dynamics. The concept of visibility dynamics accounts for some of these mechanisms, such as the bandwagon effect. When one article is cited by many subsequent publications, even more people become aware of this article. Thus, its visibility, and thereby the chances of getting even more citations, increases. This is a variant of the "Matthew effect" (Merton, 1968), stating that recognition is skewed in favor of established scientists. Similarly, when an article has received many citations, it obtains status as an authoritative paper. In turn even more authors will cite it, as appealing to existing authorities may be one reason for citing a paper (Gilbert, 1977).⁴

As indicated above, previous studies of the citation process have not provided any simple answer to the question of what citations stand for. Even now, in spite of detailed studies of referencing behavior, there is no unified theory. Nevertheless, some overall findings remain: the references have a multitude of functions in the scientific article, only a small proportion of the relevant literature is cited, and the authors have a multitude of motives for including particular studies as references. To what extent this affects the use of citations as performance indicators is still a matter of debate and is discussed below.

Validation Studies

While empirical studies have revealed a multitude of factors involved in the citation process, the issue has also been approached from another angle: by comparing citation indicators with the outcome of peer review. During recent decades, many such studies have been carried out. In the

studies, assessments by peers have been typically considered as a kind of standard to which citation indicators can be validated. The basic assumption is that there should be a correlation if citations legitimately can be used as indicators of scientific performance. The studies differ in methodology and levels of investigation, ranging from individual papers, individual researchers, research groups, and departments. In the three latter cases, a collection of publications with aggregated bibliometric measures is typically compared with peer assessment. In this way, the comparative validation is less direct by focusing on how citation indicators work at aggregated levels and not at the level of individual papers.

Some studies have analyzed grant peer reviews with the aim of assessing whether applicants that have been awarded funding were more cited than unfunded applicants (see, for example, Cabezas-Clavijo et al., 2013; Hornbostel, Bohmer, Klingsporn, Neufeld, & von Ins, 2009). However, according to a recent review, the results are ambiguous (Wouters et al., 2015). While some studies have found a positive correlation between funding and citation impact, others have questioned whether grant peer review and citation impact are correlated (Bornmann, 2011).

There are also several studies analyzing the issue with respect to peer judgments of research groups. For example, Rinia, van Leeuwen, van Vuren, and van Raan (1998) showed that various citation indicators correlated significantly with peer ratings of research programs in condensed matter physics. Aksnes and Taxt (2004) analyzed the relationship between bibliometric indicators and the outcomes of a peer review of Norwegian research groups at a mathematical and natural science faculty, reporting positive but weak correlations. Other examples include van Raan (2006) who analyzed the correlation between the h-index and several standard bibliometric indicators with the results of peer-review judgment for research groups within chemistry in the Netherlands. He found that the h-index and the normalized citation impact indicator both correlated quite well with peer judgments.

In several countries, national RAEs are carried out on a regular basis. These assessments have also enabled comparative analyses of citation indicators and peer ratings. For example, such analyses have been carried out in an Italian context (Ancaiani et al., 2015). As part of the Italian RAE, the national agency ANVUR analyzed the agreement between grades attributed to journal articles by informed peer review and by bibliometric indicators. A significant degree of concordance was found "... supporting the choice of using both techniques in order to assess the quality of Italian research institutions" (Ancaiani et al., 2015, p. 254). However, the methodological fundament for this conclusion has been contested by Baccini and De Nicolao (2016), who argue that the analysis is flawed and that informed peer review and bibliometrics do not produce similar results. As mentioned in the introduction, Abramo and D'Angelo (2011) in an article, contrasting the two approaches, also claimed

that the bibliometric is by far the preferable method in the natural and formal sciences. Other examples include Oppenheim (1997) who found strong positive correlations between citation measures and the 1992 RAE ratings for British research in genetics, anatomy, and archeology—but his conclusions were criticized by Warner (2000). Several additional studies have addressed the issue in respect to subsequent RAE assessment exercises and its successor REF (for an overview, see de Rijcke et al., 2016). The most recent example is a study comparing the outcome of REF 2014 with various metrics (Higher Education Funding Council for England, 2015). The study shows that various metrics provide significantly different outcomes from the REF peer-review process. For the field-weighted citation impact, a Spearman correlation coefficient of .28 was identified at an overall level, albeit with significant variations across fields. Moreover, there were significant decreases in correlation for more recent outputs. The study concludes that metrics cannot provide a like-for-like replacement for REF peer review. Still, the study does not analyze department-level average scores which one might argue would be more relevant with respect to the REF (cf. Traag & Waltman, 2018).

Overall, it may be concluded that most of the comparative studies seem to have found a moderately positive correspondence, but the correlations identified have been far from perfect and have varied among the studies. This means that there is so far little empirical support for claiming that citations metrics reflect the same aspects of research quality or impact as peer-review assessments. However, the extent to which the correlation is seen as sufficient depends on the context of goals of the evaluation.

There are also several problems related to the fundament for such comparative studies (Aksnes & Taxt, 2004). First, a peer evaluation may involve assessments of factors besides scientific quality or aspects that are unlikely to be mirrored through citation counts. Only when citation indicators are used in the same decision context as peer review and the two address the same dimension of the research performance can one reasonably compare them. This problem is illustrated in the comparative analysis of the REF 2014 referred to above. Here, the basis for the analysis was the peer rating of quality, consisting of different elements such as originality, significance, rigor, impact, vitality, and sustainability. Second, peer assessments may not necessarily be considered as the “truth” to which bibliometric measures should correspond—the peers may be biased or mistaken in their judgments or they may lack competence to judge (Rip, 1997). Thus, both the methodological basis for comparing peer assessments and citation indicators and the assumption that the two may be expected to correlate may be questionable. Moreover, panels increasingly are considering citation measures as part of the evaluation procedure, which means that the two cannot be considered as completely independent of each other. This relates to another issue that there is reciprocal influence which means that high citation counts may be considered as

equivalent to scientific quality. For example, according to Wouters (1999a), publishing in journals with a high impact factor has become an independent measure of scientific quality (see also Rushforth & de Rijcke, 2015). Finally, a large number of different citation measures exist and the outcome would also depend on which indicators are selected for the comparative analysis.

Citations as Indicators—Other Validity Issues

As is evident from the overview above, there is no simple answer to the question what citation indicators measure or indicate. It is clear that many limitations are attached to citations as performance measures. Besides the fundamental problems associated with the multifaceted referencing behavior of researchers, there are several more specific problems and limitations of citation indicators.

One important issue concerns the coverage of the databases applied, as well as the reference patterns. **In the social sciences and humanities, publishing in books is more common and international journals have a less prominent role.** Besides, the older literature is still important and many of the research fields have a “local” orientation (Ossenblok, Engels, & Sivertsen, 2012). **Although the literature coverage of citation databases has improved (Web of Science and Scopus), the coverage of the humanities and several social science disciplines remains limited** (Waltman, 2016). Accordingly, citation analyses may lack justification in these fields, and some countries such as Italy, which have used quantitative indicators in their national research assessments, have not included metrics in the assessments of social sciences and humanities (Ancaiani et al., 2015).

Problems related to more technical issues, such as discrepancies between target articles and cited references (misspellings of journal names, author names, errors in the reference lists, etc.), and mistakes in the indexing procedures conducted by Clarivate Analytics (previously Thomson Reuters) or Elsevier (Leydesdorff et al., 2016; Moed, 2002) may confuse citation analyses. Such errors affect in particular the accuracy of the citation counts to individual articles. A large number of more specific factors may undermine the use of citations as performance measures (see, for example, Seglen, 1997). Some of these relate to the citation process, for example, so-called “negative” citations (criticizing, correcting, and disclaiming other works), “citation circles” (groups of researchers who cite one another’s work), and extensive self-citation rates. Some of these problems have a fundamental character and are inherent in any use of citations as indicators, others may be resolved by the construction of more advanced indicators, while others again may be of less importance in practice. For example, negative citations tend to be very rare (Catalini, Lacetera, & Oettl, 2015) and self-citations can be adjusted for if needed.

However, problems and limitations of citation analysis arise differently at different levels of aggregation (Aksnes, 2005). When citations are used as indicators, aggregated levels representing larger number of papers and citations are usually analyzed. According to Welljams-Dorof (1997), this has important implications:

In general, the larger the citation data set being used, the higher the confidence level of the results. Analyses involving entire fields of research, nations, regions and large universities are virtually unaffected by the concerns and caveats about citation data . . . The confidence level at these large aggregate levels is quite high in analyses of fundamental, basic research. (p. 206)

Nevertheless, there is a lack of empirical studies confirming that this is actually the case, and possibly some of the biases is of a fundamental nature attached to all citations measures, while the effect of others may tend to level out when aggregated levels are considered.

An example of the first type of limitation relates to the skewed citation distributions. One may question whether the very highly cited papers are an order of magnitude more influential than the papers which have been less highly cited. Ideally, one wants citation indicators to measure impact in a monotonic fashion: the higher the scores, the “better” the paper (Ioannidis et al., 2016). However, according to Aksnes (2003), the skewness in the citation distribution is larger than the quality differentiation among scientific contributions might justify. This is because of the sociological and aggregational processes involved. In the beginning, an article may be cited for substantive reasons (e.g., its content has been used). Later, when the article is widely known and has received many citations, sociological mechanisms will be of increasing importance (authors citing authoritative papers, the bandwagon effect, etc.). Some papers will benefit greatly from such effects while others will not.

As described in the introduction, a large number of citation indicators exist, each with various strengths and limitations. Because of this, it has long been emphasized by bibliometricians that more than one indicator should be used in research evaluation contexts (van Raan, 1993). For example, the mean normalized citation score is size-independent and does not take into account the number of publications. According to Abramo and D’Angelo (2016), this is a major problem with this indicator because it does not truly represent productivity. The fact that citation distributions are extremely skewed also raises questions concerning the use of mean as indicator, and Bornmann and Mutz (2011) have proposed to use percentile ranks as a non-parametric alternative to the means of citation distributions for the normalization.

Dimensions of Research Quality and Citations

As shown above, the question on the relation between citations and research quality is complex and will arise

differently depending on the field analyzed, the database used, the timeframe and indicators applied, and so forth. In addition, research quality is a multidimensional concept, and in this section, we will look further into this issue.

As a starting point, we can take the three dimensions distinguished by Polanyi (1962): plausibility, originality, and scientific value.⁵ In this view, good research is based on evidence and is scientifically sound (plausibility), it provides new knowledge (originality), and it has importance for other research (scientific value). More recent studies have added societal value, that is, including importance for society as a fourth dimension of research quality (Gulbrandsen, 2000; Lamont, 2009). In many research evaluation exercises, scientific quality and societal importance/impact are seen as two independent pillars (e.g., in the U.K. REF, in the Dutch SEP, and in the most recent evaluations performed by the Research Council of Norway).

Notably, empirical studies of researchers’ conceptions of research quality have come up with a multitude of notions and aspects of quality. They span from correctness, rigor, clarity, productivity, recognition, novelty, beauty, significance, autonomy, difficulty, and relevance to ethical/sustainable research (Aksnes & Rip, 2009; Bazeley, 2010; Hemlin, 1991; Hug, Ochsner, & Daniel, 2013; Lamont, 2009; Martensson, Fors, Wallin, Zander, & Nilsson, 2016). Overall dimensions can be seen as attempts to create overall categories across such multitude of criteria and aspects.

Moreover, all assessments of research quality may be context-dependent, in terms of, for example, the time of assessment and the time/field/sector perspectives of the evaluators. Different evaluators may have different perceptions of what is significant and solid research, and what is original will by definition change over time. There may also be intrinsic tensions between the dimensions. Whereas solidity and scientific value demand some compliance with previously established norms and previous research, the most original research may conflict with this (Luukkonen, 2012; Polanyi, 1962).

In sum, whereas plausibility/soundness, scientific value, and societal value and originality seem commonly perceived key characteristics of research quality, each of these dimensions include a variety of aspects; they may be context-dependent and may also conflict with each other.

Below we discuss how citations may relate to each of these dimensions of the quality concept. Surprisingly, this topic has rarely been addressed specifically in the literature and there are few studies analyzing the issue empirically. Studies of referencing behavior have provided some findings of indirect relevance. However, from citation counts alone one cannot reveal why a specific paper is repeatedly cited by other researchers. A general methodological problem is that the multiple causes of references cannot be deduced by “travelling back” from citations. The reason for this is that the way citation indexing has developed historically leads to the loss of information about the citing context in the citation databases (Wouters, 1999b, 2014). The many different

reasons for the citations to a paper have therefore become obliterated from the record. As a result, citations cannot be sorted in those citations that do signify the perceived quality of the cited paper and those that do not.

In the following, we illustrate this further by looking at the different dimensions that together constitute the commonly used concept of “research quality.”

Solidity and Plausibility

The first dimension of the quality concept regards the plausibility, soundness, and solidity of the research. Included are virtues such as that research should be well-founded, based on scientific methods, and produce convincing results.

How citations relate to or reflect these aspects of the quality concept is complex to assess as many different dimensions need to be considered. Even when solidity and related academic virtues are aspects which are considered by peers when manuscripts are submitted to journal for publications, there are large differences when it comes to the solidity and plausibility of published studies. The literature contains numerous publications of which the solidity is poor, the results unreliable or even involving misconduct or scientific fraud (Fanelli, 2009). The latter issue has also been investigated empirically, showing that some publications which have been retracted due to fabrication and falsification of results are very highly cited, some with several hundreds of citations (Fang, Steen, & Casadevall, 2012). Moreover, a disproportionally high share of the articles retracted due to fraud were published in prestigious high impact journals. Although articles retracted due to fraud represent a very small percentage of the overall scientific literature, the problem may be increasing (Fang et al., 2012). The journal referees have apparently considered these papers as sufficiently solid to be published. More generally, there are also indications that methodological soundness and plausibility are not sufficiently emphasized in the review of manuscripts for publication (Lee, 2015). Thus, the referee system does not fully ensure the quality dimension related to solidity and plausibility, and there are no indications that high citation counts reflect solidity.

The issue may be considered from another angle: that of the reader and potential citer. One might think that in cases where the solidity or plausibility is assessed as poor, the work will not be considered as worth citing (i.e., will be neglected), and in cases where more than one study shows similar results, an author may choose to cite the study she perceives as the most solid. As a consequence, solidity/plausibility—as perceived at the time of citing—may to a certain extent be reflected in citation patterns. There is, however, little knowledge about the extent to which this actually is the case, and (as explained in “Understanding Citations” section) studies of citation behavior have identified a multitude of factors that are not per se associated with the solidity of the studies. Therefore, it seems unlikely that citations can be seen as valid indicators of the solidity of the publications.

Originality and Novelty

The second dimension, originality and novelty, derives from the fundamental demand that research should produce new knowledge. Originality may include new hypothesis, new methods, new theories and models, and new results, and may span from additions/improvements of established knowledge to radical novelty/disruption of existing research.

It seems reasonable to assume that studies with high originality or novelty will be much cited. For example, it has been argued that potential breakthrough discoveries in science can be identified on the basis of citation patterns (Winnink, Tijssen, & van Raan, 2016). Moreover, Nobel Laurates, who presumably have contributed to research of extraordinary high originality and novelty, tend to be more highly cited than the average scientists (Gingras & Wallace, 2010; Wagner, Horlings, Whetsell, Mattsson, & Nordqvist, 2015), and many have published so-called “citation classics.” Based on such observations, Garfield previously explored the possibility for using citation statistics to predict future winners (Garfield & Welljams-Dorof, 1992). At the same time, high citation counts do not necessarily imply breakthrough or Nobel class research. The extremely highly cited Lowry et al. (1951) paper on protein measurement, described above, is an interesting case in this respect. As a consequence of referencing norms, the article has probably been cited almost every time the method has been used. But according to Lowry himself, “It just happened to be a trifle better or easier or more sensitive than other methods, and of course nearly everyone measures proteins these days” (quoted in Garfield, 1979b, pp. 363-364).

Example of papers which typically would be considered to have low originality and novelty would be the so-called “replication studies.” Although such studies are important for the validation of research, for testing and demonstrating the generalizability of existing findings, they tend to be seen as “bricklaying” exercises, rather than as major contributions to the field (Everett & Earp, 2015). If the results of studies only corroborate those of previous studies, they have low novelty and are probably less likely to be cited. Many journals appear to be reluctant to publish replications because they would have a negative influence on the citation rate, the impact factor, of the journal (G. N. Martin & Clarke, 2017). However, the recent attention to the lack of replicable results in biomedical, clinical, and psychological studies (Ioannidis, 2005) may lead to a higher social status of replications studies.

The above considerations show that there is no simple relationship between originality or novelty and citations. Studies with high originality may include both major scientific advances and minor contributions. In the latter case, articles may not be cited because their research question is a “dead end” which means that it does not function as a positive basis for further work—despite being novel or original in approach. This brings us to the next dimension of the research quality, scientific value.

Scientific Value

Scholarly or scientific significance may include relevance to previous as well as future research—cumulatively as well as the opening new research fields. Assessments of the importance of the research may depend on the generalizability of the results and the size of, and general interest in, the research field/question.

Scientific value and significance are dimensions of the quality concept to which some citations may most directly relate. This is commonly argued as follows. When a scientist refers to a paper, it has been useful or relevant in some way for the present research or for the writing of the publication. Thus, frequently cited articles may be assumed to have been more useful than publications that are hardly cited or not at all, and possibly be more useful and thus important in their own right (Aksnes, 2005). This means that the number of citations may be considered as a measure of the article's usefulness, impact, or influence on other research. The same reasoning can be used for aggregated levels of articles. This is the typical way of justifying the use of citations as performance indicator. However, as discussed in "Understanding Citations" section, citations have both intellectual and social functions. In recent times, the relationship between scholarly quality and citations has become more complicated as researchers have become aware of the need to increase their visibility. This has become especially urgent as research funding has become scarce and the competition for resources has sharpened. In addition, since the use of citation indicators as performance indicators, researchers are aware that their references may influence the careers of the researchers they cite. High numbers of citations to a particular research group or individual researcher may thus be the result of a strong visibility strategy or of direct or indirect "citation gaming" (Biagioli & Lippmann, in press). Although strategies to strategically cite are not by definition questionable research practices (but some of them would certainly qualify as such), these processes do undermine the validity of the citation as an indicator of scholarly quality.

In 1983, B. R. Martin and Irvine described the conceptual difference between quality and impact in this way: "'Quality' is a property of the publication and the research described in it. It describes how well the research has been done, whether it is free from obvious 'error' . . . how original the conclusions are, and so on" (p. 70). The impact of a publication, on the contrary, is defined as the "*actual* influence on surrounding research activities at a given time." In the view of B. R. Martin and Irvine, it is the impact of a publication that most closely is related to the concept of scientific progress—a publication causing a great impact represents a major contribution to knowledge at the time it is published. Using these definitions, it is also evident that impact would be a more adequate interpretation of citations than quality. As an example, even a "mistaken" publication can have a large impact by stimulating further research. Similarly, a publication by a

recognized scientist may be more visible and therefore have more impact, earning more citations, even if its quality (in terms of originality and solidity) is no greater than those by lesser known researchers (B. R. Martin, 1996). Impact is the most commonly used concept for what citations reflect, although other concepts such as influence, importance, significance, and utility occasionally also are used (Moed, 2005). However, the use of *impact* as the most appropriate concept has usually been justified by theoretical considerations, and there are few attempts to address the issue empirically or relate it to previous findings on citation behavior. Some have attempted to resolve this issue by using the combined concept *citation impact*, as this expresses the methodology used to measure impact (Moed, 2005). According to Waltman, van Eck, and Wouters (2013), citation impact should be distinguished from scientific impact, as an influential researcher sometimes has a lower performance in terms of highly cited publications than some of their less influential colleagues.

Societal Value and Relevance

This dimension of the quality concept may include any kind of extra-scientific relevance, for example, relevance to education, health, wealth, or the environment. In many settings, research with a value outside science will be higher valued, and social relevance and broader impacts are often part of funding agencies' review criteria for research grants (Langfeldt & Scordato, 2015).

Societal relevance is often considered to be something which is much harder to measure than scientific relevance or impact (B. R. Martin, 2011). There seems to be a widespread assumption that this issue cannot be adequately assessed through standard citation indicators, and in recent years, increasing attention has been devoted to developing methodologies for assessing and measuring societal relevance and impact (Bornmann, 2012, 2013).

For a long time, citation analyses have been applied in patent studies (Meyer, 2000). Through analyses of patents citations to scientific publications, knowledge has been obtained on the interaction and impact of science on technology. Thus, these studies have yielded information on a particular type of societal relevance and impact: the technological (van Raan, 2017). Still, a basic limitation is that many innovations are not patented and patents are not suitable to assess societal relevance or impact in a broader context. Only a very small minority of the publications indexed in the Web of Science or Scopus databases are actually cited by patents (van Raan, 2017).

A more general reason why societal relevance is difficult to assess through citation counts is that the literature indexed in Web of Science and Scopus consists mostly of academic and scholarly publications.⁶ While citations may reflect intra-scientific use, use and applications that take place along other dimensions are far less likely to be captured by citation

counts in such journals. For example, Hanney et al. (2006) showed that some diabetes papers which were assessed as having had an important impact on clinical practice did not receive many citations.

Citation indicators are also often considered to have important limitations in applied areas. For example, le Pair (1995) has emphasized, “In technology or practicable research bibliometrics is an insufficient means of evaluation. It may help a little, but just as often it may lead to erroneous conclusions” (p. 18). Similarly, research of mainly national or local interest may often be poorly cited by the literature published in international academic journals.

Nevertheless, it is clear that scientific contributions with large societal relevance may also be highly cited. For example, Edward C. Prescott and Finn E. Kydland received the Nobel Memorial Prize in Economics in 2004 for two papers which profoundly influenced the practice of economic policy in general, and monetary policy in particular (Dymond, 2015). These papers are very highly cited also in the academic literature. Similarly, in 1994, the Scandinavian Simvastatin Survival Study (4S) provided the first unequivocal evidence that lowering low-density lipoprotein (LDL) cholesterol via statin treatment reduces cardiovascular events and overall mortality (Pedersen et al., 1994). This paper is now cited more than 7,700 times in the Web of Science database. Simvastatin was developed by Merck & Co. and came into medical use in 1992 and has had major impact on human health (Li, 2009). Prior to losing its patent protection, simvastatin was Merck’s largest-selling drug and second-largest-selling cholesterol lowering drug in the world. Despite these and numerous similar examples, it is not possible to identify societal relevance from citation counts per se, and uncited or little cited publication may have contributed with results of great societal relevance.

As described above, there is currently much interest toward developing alternative indicators that could capture these aspects of scientific activities better, which would be undervalued when using traditional citation-based indicators. This includes altmetrics using data from social media sources (Weller, 2015), and the development of models for analyzing the impact of research, such as the “Payback Framework” (Donovan & Hanney, 2011). New forms of citation analyses have also been developed for analyzing societal impact of research. For example, the impact of research on health care has been investigated using data on publications cited in clinical guidelines (Grant, Cottrell, Cluzeau, & Fawcett, 2000; Lewison & Sullivan, 2008). Similarly, new methods and templates for classification of citations have been introduced for assessing how research findings are translated and used in clinical practice (Jones & Hanney, 2016).

Concluding Remarks

The use of citation indicators in research evaluation contexts has increased in recent years, as described previously. The view generally held among experts within bibliometrics

seem to be that citations represent a good but not perfect impact measure. However, considering the various limitations attached to citations as performance measures, most bibliometricians have argued that a bibliometric analysis should not function as a substitute for a peer review (Moed, 2005). At the same time, there are also various limitations and shortcomings with peer assessments (Chubin & Hackett, 1990). For example, **human judgment is subjective and the opinions of experts may be influenced by lack of knowledge and limited cognitive horizons** (Lee, Sugimoto, Zhang, & Cronin, 2013; van Raan, 2000). Moreover, peer reviews are expensive and slow.⁷

On this basis, it is often argued that bibliometric analysis can counterbalance shortcomings and mistakes in the peers’ judgments (Aksnes, 2005). Thus, a bibliometric study should be considered as complementary to a peer evaluation (Council of Canadian Academies, 2012). According to Aksnes and Taxt (2004), such a combination of methods would have improved the reliability of evaluations carried out in Norway. In cases with large discrepancies between the peers’ qualitative judgments and the bibliometric performance measures, the evaluation committee should investigate the reasons for these deviations. Then, they might find that their own assessments are mistaken or that the bibliometric measures did not reflect the unit’s performance (van Raan, 1996).

In the REF 2014, citation analyses were carried out for 11 of the 36 field-delineated subpanels, mostly in the life- and physical-science areas (Wilsdon et al., 2015). In the report on the role of metrics in research assessment and management, it is recommended that “quantitative data—particularly around published outputs—continue to have a place in informing peer-review judgments of research quality. This approach has been used successfully in REF2014, and we recommend that it be continued and enhanced in future exercises” (Wilsdon et al., 2015). However, at the same time, it is warned,

Bibliometricians generally see citation rates as a proxy measure of academic impact or of impact on the relevant academic communities. But this is only one of the dimensions of academic quality. Quality needs to be seen as a multidimensional concept that cannot be captured by any one indicator, and which dimension of quality should be prioritised may vary by field and mission.

As is evident from the discussion of this paper, this is an important point, as citations are not able to capture all aspects of the quality concept. Hence, an increased use of citation indicators in research evaluation and funding may imply less attention to these other research quality dimensions, such as solidity/plausibility, originality, and societal value.

Since the introduction of citation-based indicators in research evaluation contexts several decades ago, there have often been controversies surrounding the applications (Wouters, 1999b). **The use of bibliometric indicators for evaluation purposes is sometimes met by opposition by scientific communities.** For example, researchers are concerned

about possible lack of fairness, particularly if the evaluations have consequences for research funding. Evaluations which are critical or negative often generate protests, although this applies to all evaluations regardless of methods applied (Luukkonen, 1997a). At the same time, others have welcomed the use of citation indicators. The recent report on the use of metrics in the REF also shows that there is huge variation in the viewpoints within the scholarly and scientific communities (Wilsdon et al., 2015).

There are no indications that the use of citations as performance indicators will subside in the future. Against this background, sensible use of indicators is important. Citation indicators may easily be misused or applied in contexts where they lack justification or validity. There is a growing concern about this issue, as well as on potential negative impact of research metrics on the scientific community. This is exemplified by the publication of the Leiden manifesto containing 10 principles for the measurement of research performance (Hicks et al., 2015) and the San Francisco Declaration on Research Assessment (DORA) which intends to prevent the practice of using the JIF “. . . as a surrogate measure of the quality of individual research articles, to assess an individual scientist’s contributions, or in hiring, promotion, or funding decisions,” (p. 869) (Cagan, 2013).

We conclude that citations reflect—with important limitations—aspects related to scientific impact and relevance, but there is no evidence that citations reflect other key dimensions of research quality. There is no obvious road to better handle the tension between administrative needs for simple measures and more easy evaluation methods and researchers’ request for fair and comprehensive assessments of scientific quality. Citation-based indicators cannot provide sufficiently nuanced or robust measures of quality when used in isolation. At the same time, there are also problems with the peer-review system. However, the viewpoint described in the introduction that bibliometric assessment is superior compared to the traditional peer-review method is not justified in our opinion. Peer reviews are applied in many different contexts, of which peer assessments of manuscripts submitted to journals and publishers probably is the most fundamental one. For such assessment, citation indicators are hardly of any relevance. More generally, citation indicators seem of little help in the evaluation of the solidity/plausibility, originality, and societal value of research.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research was funded by the Research Council of Norway (RCN), grant number 256223 (the R-QUEST centre).

Notes

1. For example, they claim, “Overall, there are reasons to support bibliometrics-based review beyond cost considerations. Even simple metrics can perform well at identifying quality for some fields, while providing cost effective and transparent review. Peer review does not appear to be a guarantor of quality . . .” (Regibeau & Rockett, 2016).
2. The article is partly based on literature reviews first conducted for one of the author’s doctoral dissertation (Aksnes, 2005), which have been combined and extended with more recent contributions. Some text passages from this dissertation have been adapted and incorporated into the present article.
3. As an example, Garfield (1979a) early emphasized that “Instead of directly comparing the citation counts of, say, a mathematician against that of a biochemist, both should be ranked with their peers, and the comparison should be made between rankings” (p. 367).
4. According to Small, it may be assumed that highly cited papers represent the key concepts, methods, or experiments in a field. Frequently cited papers have been viewed as “exemplars” (using Thomas Kuhn’s terminology), whereby papers are cited because they represent a classical study, a “concept” marker (Small, 1978), or show how a particular line of research is carried out.
5. Notably, Polanyi used the term “scientific merit,” not “quality.” Quality may be a broader term, encompassing more aspects than merit. Still, we believe Polanyi addressed the same issues as those relevant to our discussion of research quality and citation indicators.
6. However, according to the web page of Scopus (<https://www.elsevier.com/solutions/scopus/content>), more than 300 trade journals reaching a specific industry, trade, or type of business have been selected for Scopus coverage.
7. For example, Eyre-Walker and Stoletzki (2013) argued, “In particular subjective peer review is error prone, biased, and expensive; we must therefore question whether using peer review in exercises such as the research assessment exercise (RAE) and the Research Excellence Framework (REF) is worth the huge amount of resources spent on them” (p. e1001675).

ORCID iD

Dag W. Aksnes  <https://orcid.org/0000-0002-1519-195X>

References

- 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., & D’Angelo, C. A. (2016). A farewell to the MNCS and like size-independent indicators. *Journal of Informetrics*, 10, 646-651. doi:10.1016/j.joi.2016.04.006
- Aksnes, D. W. (2003). Characteristics of highly cited papers. *Research Evaluation*, 12, 159-170.
- Aksnes, D. W. (2005). *Citations and their use as indicators in science policy: Studies of validity and applicability issues with a particular focus on highly cited papers* (Doctoral thesis). University of Twente, Enschede, The Netherlands.

- Aksnes, D. W., & Rip, A. (2009). Researchers' perceptions of citations. *Research Policy*, 38, 895-905. doi:10.1016/j.respol.2009.02.001
- Aksnes, D. W., & Taxt, R. E. (2004). Peer reviews and bibliometric indicators: A comparative study at a Norwegian university. *Research Evaluation*, 13, 33-41. doi:10.3152/147154404781776563
- 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, 273-289. doi:10.1016/j.joi.2009.04.001
- Amsterdamska, O., & Leydesdorff, L. (1989). Citations: Indicators of significance? *Scientometrics*, 15, 449-471.
- Ancaiani, A., Anfossi, A. F., Barbara, A., Benedetto, S., Blasi, B., Carletti, V., . . . Sileoni, S. (2015). Evaluating scientific research in Italy: The 2004-10 research evaluation exercise. *Research Evaluation*, 24, 242-255. doi:10.1093/reseval/rvv008
- Baccini, A., & De Nicolao, G. (2016). Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise. *Scientometrics*, 108, 1651-1671. doi:10.1007/s11192-016-1929-y
- Baumgartner, S. E., & Leydesdorff, L. (2014). Group-Based Trajectory Modeling (GBTM) of citations in scholarly literature: Dynamic qualities of "transient" and "sticky knowledge claims." *Journal of the Association for Information Science and Technology*, 65, 797-811. doi:10.1002/asi.23009
- Bazeley, P. (2010). Conceptualising research performance. *Studies in Higher Education*, 35, 889-903. doi:10.1080/03075070903348404
- Biagioli, M., & Lippman, A. (in press.) *Gaming Metrics. Beyond Publish or Perish: Metrics and the new Ecologies of Academic Misconduct*. MIT Press
- Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology*, 45, 199-245.
- Bornmann, L. (2012). Measuring the societal impact of research. *EMBO Reports*, 13, 673-676. doi:10.1038/embor.2012.99
- Bornmann, L. (2013). What is societal impact of research and how can it be assessed? A literature survey. *Journal of the American Society for Information Science and Technology*, 64, 217-233. doi:10.1002/asi.22803
- Bornmann, L., & Daniel, H. D. (2007). What do we know about the h index? *Journal of the American Society for Information Science and Technology*, 58, 1381-1385. doi:10.1002/asi.20609
- Bornmann, L., & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64, 45-80. doi:10.1108/00220410810844150
- Bornmann, L., Marx, W., Gasparyan, A. Y., & Kitas, G. (2012). Diversity, value and limitations of the journal impact factor and alternative metrics. *Rheumatology International*, 32, 1861-1867. doi:10.1007/s00296-011-2276-1
- 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, 228-230. doi:10.1016/j.joi.2010.10.009
- Cabezas-Clavijo, A., Robinson-Garcia, N., Escabias, M., & Jimenez-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
- Cagan, R. (2013). The San Francisco Declaration on Research Assessment. *Disease Models & Mechanisms*, 6, 869-870. doi:10.1242/dmm.012955
- Camacho-Minano, M. D. M., & Nunez-Nickel, M. (2009). The multilayered nature of reference selection. *Journal of the American Society for Information Science and Technology*, 60, 754-777. doi:10.1002/asi.21018
- Carlsson, H. (2009). Allocation of research funds using bibliometric indicators—Asset and challenge to Swedish higher education sector. *InfoTrend*, 64(4), 82-88.
- Catalini, C., Lacetera, N., & Oettl, A. (2015). The incidence and role of negative citations in science. *Proceedings of the National Academy of Sciences of the United States of America*, 112, 13823-13826. doi:10.1073/pnas.1502280112
- Chubin, D. E., & Hackett, E. J. (1990). *Peerless science: Peer review and U.S. science policy*. Albany: State University of New York Press.
- Cole, J. R., & Cole, S. (1973). *Social stratification in science*. Chicago, IL: The University of Chicago Press.
- Cole, S. (1992). *Making science: Between nature and society*. London, England: Harvard University Press.
- Cole, S. (2000). The role of journals in the growth of scientific knowledge. In B. Cronin & H. B. Atkins (Eds.), *The web of knowledge: A Festschrift in honor of Eugene Garfield* (pp. 109-142). Medford, NJ: American Society for Information Science.
- Council of Canadian Academies. (2012). *Informing research choices: Indicators and judgment: The expert panel on science performance and research funding*. Retrieved from <https://www.scienceadvice.ca/reports/informing-research-choices-indicators-and-judgment/>
- Cozzens, S. E. (1989). What do citations count? The rhetoric-first model. *Scientometrics*, 15, 437-447.
- Cronin, B. (1984). *The citation process: The role and significance of citations in scientific communication*. London, England: Taylor Graham.
- de Rijcke, S., Wouters, P. F., Rushforth, A. D., Franssen, T. P., & Hammarfelt, B. (2016). Evaluation practices and effects of indicator use—A literature review. *Research Evaluation*, 25, 161-169. doi:10.1093/reseval/rvv038
- Donovan, C., & Hanney, S. (2011). The "payback framework" explained. *Research Evaluation*, 20, 181-183. doi:10.3152/095820211x13118583635756
- Durieux, V., & Gevenois, P. A. (2010). Bibliometric indicators: Quality measurements of scientific publication. *Radiology*, 255, 342-351. doi:10.1148/radiol.09090626
- Dymond, L. H. (2015). *A recent history of recognized economic thought—Contributions of the Nobel laureates to economic science*. Lulu Publishing Services.
- Everett, J. A. C., & Earp, B. D. (2015). A tragedy of the (academic) commons: Interpreting the replication crisis in psychology as a social dilemma for early-career researchers. *Frontiers in Psychology*, 6, 1152. doi:10.3389/fpsyg.2015.01152
- 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
- Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE*, 4(5), e5738. doi:10.1371/journal.pone.0005738

- Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences of the United States of America*, 109, 17028-17033. doi:10.1073/pnas.1212247109
- Garfield, E. (1977). Can citation indexing be automated? In *Essay of an information scientist* (Vol. 1, pp. 84-90). Philadelphia, PA: ISI Press.
- Garfield, E. (1979a). *Citation indexing—Its theory and application in science, technology and humanities*. New York, NY: John Wiley.
- Garfield, E. (1979b). Is citation analysis a legitimate evaluation tool? *Scientometrics*, 1, 359-375.
- Garfield, E. (1997). Validation of citation analysis. *Journal of the American Society for Information Science*, 48, 962-963.
- Garfield, E., & Welljams-Dorof, A. (1992). Of Nobel class: A citation perspective on high impact research authors. *Theoretical Medicine*, 13, 117-135. doi:10.1007/bf02163625
- Gilbert, N. G. (1977). Referencing as persuasion. *Social Studies of Science*, 7, 113-122.
- Gingras, Y., & Wallace, M. L. (2010). Why it has become more difficult to predict Nobel Prize winners: A bibliometric analysis of nominees and winners of the chemistry and physics prizes (1901-2007). *Scientometrics*, 82, 401-412. doi:10.1007/s11192-009-0035-9
- Glaser, J., & Laudel, G. (2001). Integrating scientometric indicators into sociological studies: Methodical and methodological problems. *Scientometrics*, 52, 411-434. doi:10.1023/a:1014243832084
- Grant, J., Cottrell, R., Cluzeau, F., & Fawcett, G. (2000). Evaluating "payback" on biomedical research from papers cited in clinical guidelines: Applied bibliometric study. *British Medical Journal*, 320, Article 1107. doi:10.1136/bmj.320.7242.1107
- Gulbrandsen, J. M. (2000). *Research quality and organisational factors: An investigation of the relationship* (Doctoral thesis). Norwegian University of Science and Technology, Trondheim.
- Hanney, S. R., Home, P. D., Frame, I., Grant, J., Green, P., & Buxton, M. J. (2006). Identifying the impact of diabetes research. *Diabetic Medicine*, 23, 176-184. doi:10.1111/j.1464-5491.2005.01753.x
- Harzing, A. W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison. *Scientometrics*, 106, 787-804. doi:10.1007/s11192-015-1798-9
- Higher Education Funding Council for England. (2015). *The metric tide: Correlation analysis of REF2014 scores and metrics: Supplementary report II to the independent review of the role of metrics in research assessment and management*. Retrieved from www.dcsience.net/2015_metrictideS2.pdf
- Hellqvist, B. (2010). Referencing in the humanities and its implications for citation analysis. *Journal of the American Society for Information Science and Technology*, 61, 310-318. doi:10.1002/asi.21256
- Hemlin, S. (1991). *Quality in science: Researchers' conceptions and judgments* (Doctoral thesis). University of Göteborg, Sweden.
- Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, 520, 429-431.
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 16569-16572. doi:10.1073/pnas.0507655102
- Holden, G., Rosenberg, G., & Barker, K. (2005). Bibliometrics: A potential decision making aid in hiring, reappointment, tenure and promotion decisions. In G. Holden, G. Rosenberg, & K. Barker (Eds.), *Bibliometrics in social work* (pp. 67-92). New York, NY: Routledge.
- Hornbostel, S., Bohmer, S., Klingsporn, B., Neufeld, J., & von Ins, M. (2009). Funding of young scientist and scientific excellence. *Scientometrics*, 79, 171-190. doi:10.1007/s11192-009-0411-5
- 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, 369-383. doi:10.1093/reseval/rvt008
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2, 696-701. doi:10.1371/journal.pmed.0020124
- Ioannidis, J. P. A., Boyack, K., & Wouters, P. F. (2016). Citation metrics: A primer on how (not) to normalize. *PLoS Biology*, 14(9), e1002542. doi:10.1371/journal.pbio.1002542
- Jones, T. H., & Hanney, S. (2016). Tracing the indirect societal impacts of biomedical research: Development and piloting of a technique based on citations. *Scientometrics*, 107, 975-1003. doi:10.1007/s11192-016-1895-4
- Lamont, M. (2009). *How professors think: Inside the curious world of academic judgment*. Cambridge, MA: Harvard University Press.
- Langfeldt, L., Benner, M., Sivertsen, G., Kristiansen, E. H., Aksnes, D. W., Borlaug, S. B., . . . Pelkonen, A. (2015). Excellence and growth dynamics: A comparative study of the Matthew effect. *Science and Public Policy*, 42, 661-675. doi:10.1093/scipol/scu083
- Langfeldt, L., & Scordato, L. (2015, April). *Assessing the broader impacts of research: A review of methods and practices* (NIFU Working Paper 8/2015). Oslo.
- Lariviere, V., & Gingras, Y. (2011). Averages of ratios vs. ratios of averages: An empirical analysis of four levels of aggregation. *Journal of Informetrics*, 5, 392-399. doi:10.1016/j.joi.2011.02.001
- Law, J. (1986). The heterogeneity of texts. In M. Callon, J. Law, & A. Rip (Eds.), *Mapping the dynamics of science and technology* (pp. 67-83). London, England: Macmillan.
- Lee, C. J. (2015). Commensuration bias in peer review. *Philosophy of Science*, 82, 1272-1283. doi:10.1086/683652
- 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, 2-17. doi:10.1002/asi.22784
- le Pair, C. (1995). Formal evaluation methods: Their utility and limitations. *International Forum on Information and Documentation*, 20(4), 16-24.
- Levitt, J. M., & Thelwall, M. (2011). A combined bibliometric indicator to predict article impact. *Information Processing & Management*, 47, 300-308. doi:10.1016/j.ipm.2010.09.005
- Lewison, G., & Sullivan, R. (2008). The impact of cancer research: How publications influence UK cancer clinical guidelines. *British Journal of Cancer*, 98, 1944-1950. doi:10.1038/sj.bjc.6604405

- Leydesdorff, L. (1989). Words and co-words as indicators of intellectual organization. *Research Policy*, 18, 209-223.
- Leydesdorff, L., Wouters, P., & Bornmann, L. (2016). Professional and citizen bibliometrics: Complementarities and ambivalences in the development and use of indicators—A state-of-the-art report. *Scientometrics*, 109, 2129-2150.
- Li, J. J. (2009). *Triumph of the heart: The story of statins*. New York, NY: Oxford University Press.
- Liu, M. X. (1993). Progress in documentation the complexities of citation practice: A review of citation studies. *Journal of Documentation*, 49, 370-408. doi:10.1108/eb026920
- Lowry, O. H., Rosebrough, N. J., Farr, A. L., & Randal, R. J. (1951). Protein measurement with the Folin phenol reagent. *Journal of Biological Chemistry*, 193, 265-275.
- Luukkonen, T. (1990). *Citations in the rhetorical, reward, and communication systems of science* (Doctoral thesis). University of Tampere, Finland.
- Luukkonen, T. (1997a). Quantitative techniques in evaluation in Western Europe. In M. S. Finkel & J. Cave (Eds.), *Evaluating science and scientists: An East-West dialogue on research evaluation in post-communist Europe* (pp. 115-131). Budapest, Hungary: Central European University Press.
- Luukkonen, T. (1997b). Why has Latour's theory of citations been ignored by the bibliometric community? Discussion of sociological interpretations of citation analysis. *Scientometrics*, 38, 27-37.
- Luukkonen, T. (2012). Conservatism and risk-taking in peer review: Emerging ERC practices. *Research Evaluation*, 21, 48-60. doi:10.1093/reseval/rvs001
- MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. *Journal of the American Society for Information Science*, 40, 342-349.
- MacRoberts, M. H., & MacRoberts, B. R. (1996). Problems of citation analysis. *Scientometrics*, 36, 435-444.
- Martensson, P., Fors, U., Wallin, S. B., Zander, U., & Nilsson, G. H. (2016). Evaluating research: A multidisciplinary approach to assessing research practice and quality. *Research Policy*, 45, 593-603. doi:10.1016/j.respol.2015.11.009
- Martin, B. R. (1996). The use of multiple indicators in the assessment of basic research. *Scientometrics*, 36, 343-362.
- Martin, B. R. (2011). The research excellence framework and the "impact agenda": Are we creating a Frankenstein monster? *Research Evaluation*, 20, 247-254. doi:10.3152/095820211x13118583635693
- Martin, B. R., & Irvine, J. (1983). Assessing basic research: Some partial indicators of scientific progress in radio astronomy. *Research Policy*, 12, 61-90.
- Martin, G. N., & Clarke, R. M. (2017). Are psychology journals anti-replication? A snapshot of editorial practices. *Frontiers in Psychology*, 8, 523.
- Marx, W., & Bornmann, L. (2015). On the causes of subject-specific citation rates in Web of Science. *Scientometrics*, 102, 1823-1827. doi:10.1007/s11192-014-1499-9
- Merton, R. K. (1968). The Matthew effect in science. *Science*, 159, 56-63.
- Merton, R. K. (1979). Foreword. In E. Garfield (Ed.), *Citation indexing—Its theory and application in science, technology, and humanities* (pp. v-ix). New York, NY: John Wiley.
- Meyer, M. (2000). Does science push technology? Patents citing scientific literature. *Research Policy*, 29, 409-434. doi:10.1016/s0048-7333(99)00040-2
- Moed, H. F. (2002). The impact-factors debate: The ISI's uses and limits. *Nature*, 415, 731-732.
- Moed, H. F. (2005). *Citation analysis in research evaluation*. Dordrecht, The Netherlands: Springer.
- Nicolaisen, J. (2007). Citation analysis. *Annual Review of Information Science and Technology*, 41, 609-641. doi:10.1002/aris.2007.1440410120
- 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.
- Ophof, T., & Leydesdorff, L. (2010). Caveats for the journal and field normalizations in the CWTS ("Leiden") evaluations of research performance. *Journal of Informetrics*, 4, 423-430. doi:10.1016/j.joi.2010.02.003
- Ossenblok, T. L. B., Engels, T. C. E., & 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, 280-290. doi:10.1093/reseval/rvs019
- Osterloh, M., & Frey, B. S. (2015). Ranking games. *Evaluation Review*, 39, 102-129. doi:10.1177/0193841x14524957
- Pedersen, T. R., Kjekshus, J., Berg, K., Haghefelt, T., Faergeman, O., Thorgeirsson, G., . . . Grundstrom, I. (1994). Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: The Scandinavian Simvastatin Survival Study (4S). *The Lancet*, 344, 1383-1389.
- Piro, F. N., & Sivertsen, G. (2016). How can differences in international university rankings be explained? *Scientometrics*, 109, 2263-2278.
- Polanyi, M. (1962). The republic of science: Its political and economic theory. *Minerva*, 1, 54-73.
- Price, D. J. d. S. (1965). Networks of scientific papers. *Science*, 149, 510-515.
- Regibeau, P., & Rockett, K. E. (2016). *Research assessment design and the role of bibliometrics*. Retrieved from <http://voxeu.org/article/using-bibliometrics-gauge-research-quality>
- Rinia, E. J., van Leeuwen, T. N., van Vuren, H. G., & van Raan, A. F. J. (1998). Comparative analysis of a set of bibliometric indicators and central peer review criteria: Evaluation of condensed matter physics in the Netherlands. *Research Policy*, 27, 95-107.
- Rip, A. (1997). Qualitative conditions of scientometrics: The new challenges. *Scientometrics*, 38, 7-26.
- Rushforth, A., & de Rijcke, S. (2015). Accounting for impact? The journal impact factor and the making of biomedical research in the Netherlands. *Minerva*, 53, 117-139. doi:10.1007/s11024-015-9274-5
- Schubert, A., & Braun, T. (1986). Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics*, 9, 281-291.
- Schubert, A., Glänzel, W., & Braun, T. (1987). World flash on basic research: Subject field characteristic citation scores and scales for assessing research performance. *Scientometrics*, 12, 267-292.

- Seglen, P. O. (1989). From bad to worse: Evaluation by journal impact. *Trends in Biochemical Sciences*, 14, 326-327.
- Seglen, P. O. (1997). Citations and journal impact factors: Questionable indicators of research quality. *Allergy*, 52, 1050-1056.
- Seglen, P. O. (1998). Citation rates and journal impact factors are not suitable for evaluation of research. *Acta Orthopaedica Scandinavica*, 69, 224-229. doi:10.3109/17453679809000920
- Small, H. G. (1978). Cited documents as concept symbols. *Social Studies of Science*, 8, 327-340.
- Small, H. G. (1982). Citation context analysis. In B. Dervin & M. Voigt (Eds.), *Progress in communication sciences* (Vol. 3, pp. 287-310). Norwood, NJ: Ablex.
- Smith, L. C. (1981). Citation analysis. *Library Trends*, 30, 83-106.
- Thornley, C., Watkinson, A., Nicholas, D., Volentine, R., Jamali, H. R., Herman, E., . . . Tenopir, C. (2015). The role of trust and authority in the citation behaviour of researchers. *Information Research: An International Electronic Journal*, 20(3), 1-17.
- Traag, V., & Waltman, L. (2018). *Systematic analysis of agreement between metrics and peer review in the UK REF*. arXiv preprint arXiv:1808.03491.
- Van Noorden, R., Maher, B., & Nuzzo, R. (2014). The top 100 papers. *Nature*, 514, 550-553.
- van Raan, A. F. J. (1993). Advanced bibliometric methods to assess research performance and scientific development: Basic principles and recent practical applications. *Research Evaluation*, 3, 151-166.
- van Raan, A. F. J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics*, 36, 397-420.
- van Raan, A. F. J. (2000). The Pandora's box of citation analysis: Measuring scientific excellence—The last evil? In B. Cronin & H. B. Atkins (Eds.), *The web of knowledge: A Festschrift in honor of Eugene Garfield* (pp. 301-319). Medford, NJ: American Society for Information Science.
- van Raan, A. F. J. (2004). Measuring science: Capita selecta of current main issues. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research* (pp. 19-50). Dordrecht: Springer.
- van Raan, A. F. J. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, 67, 491-502. doi:10.1556/Scient.67.2006.3.10
- van Raan, A. F. J. (2017). Patent citations analysis and its value in research evaluation: A review and a new approach to map technology-relevant research. *Journal of Data and Information Science*, 2, 13-50.
- Vinkler, P. (2010). *The evaluation of research by scientometric indicators*. Oxford, UK: Chandos Publishing.
- Voos, H., & Dagaev, K. S. (1976). Are all citations equal? Or, did we op. cit. your idem? *The Journal of Academic Librarianship*, 1(6), 19-21.
- Wagner, C. S., Horlings, E., Whetsell, T. A., Mattsson, P., & Nordqvist, K. (2015). Do Nobel laureates create prize-winning networks? An analysis of collaborative research in physiology or medicine. *PLoS ONE*, 10(7), e0134164. doi:10.1371/journal.pone.0134164
- Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, 10, 365-391. doi:10.1016/j.joi.2016.02.007
- Waltman, L., & Schreiber, M. (2013). On the calculation of percentile-based bibliometric indicators. *Journal of the American Society for Information Science and Technology*, 64, 372-379. doi:10.1002/asi.22775
- Waltman, L., & van Eck, N. J. (2013). Source normalized indicators of citation impact: An overview of different approaches and an empirical comparison. *Scientometrics*, 96, 699-716. doi:10.1007/s11192-012-0913-4
- Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011a). Towards a new crown indicator: An empirical analysis. *Scientometrics*, 87, 467-481. doi:10.1007/s11192-011-0354-5
- Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011b). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, 5, 37-47. doi:10.1016/j.joi.2010.08.001
- Waltman, L., van Eck, N. J., & Wouters, P. (2013). Counting publications and citations: Is more always better? *Journal of Informetrics*, 7, 635-641. doi:10.1016/j.joi.2013.04.001
- Warner, J. (2000). A critical review of the application of citation studies to the research assessment exercises. *Journal of Information Science*, 26, 453-460.
- Weingart, P. (2004). Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics*, 62, 117-131.
- Weller, K. (2015). Social media and altmetrics: An overview of current alternative approaches to measuring scholarly impact. In I. M. Welp, J. Wollersheim, S. Ringelhan, & M. Osterloh (Eds.), *Incentives and performance: Governance of research organizations* (pp. 261-279). Cambridge, UK: Springer.
- Welljams-Dorof, A. (1997). Quantitative citation data as indicators in science evaluations: A primer on their appropriate use. In M. S. Frankel & J. Cave (Eds.), *Evaluating science and scientists: An East-West dialogue on research evaluation in post-communist Europe* (pp. 202-211). Budapest, Hungary: Central European University Press.
- Willett, P. (2013). Readers' perceptions of authors' citation behaviour. *Journal of Documentation*, 69, 145-156. doi:10.1108/00220411311295360
- Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., . . . Johnson, B. (2015). *The metric tide: Report of the independent review of the role of metrics in research assessment and management*. Retrieved from <http://www.hefce.ac.uk/pubs/rereports/year/2015/metrictide/>
- Winnink, J. J., Tijssen, R. J. W., & van Raan, A. F. J. (2016). Theory-changing breakthroughs in science: The impact of research teamwork on scientific discoveries. *Journal of the Association for Information Science and Technology*, 67, 1210-1223. doi:10.1002/asi.23505
- Wouters, P. (1999a). Beyond the Holy Grail: From citation theory to indicator theories. *Scientometrics*, 44, 561-580. doi:10.1007/bf02458496
- Wouters, P. (1999b). *The citation culture* (Doctoral thesis). University of Amsterdam, The Netherlands. Retrieved from <http://garfield.library.upenn.edu/wouters/wouters.pdf>
- Wouters, P. (2014). The citation: From culture to infrastructure. In B. Cronin & C. R. Sugimoto (Eds.), *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly performance* (pp. 47-66). Cambridge: MIT Press.

- Wouters, P. (in press). Quality and impact in research evaluation. In S. Kuhlmann, D. Simon, & J. Stamm (Eds.), *Handbook of science policy studies*. Springer.
- Wouters, P., Thelwall, M., Kousha, K., Waltman, L., de Rijcke, S., Rushforth, A., & Franssen, T. (2015). *The metric tide: Literature review: Supplementary report I to the independent review of the role of metrics in research assessment and management*. Retrieved from <http://www.hefce.ac.uk/pubs/rereports/year/2015/metrictide/>

Author Biographies

Dag W. Aksnes is research professor at the Nordic Institute for studies in Innovation, Research and Education (NIFU) and

affiliated with the Centre for Research Quality and Policy Impact Studies (R-QUEST). Aksnes' research covers various topics within the field of bibliometrics, such as studies of citations, citation analyses and assessments of research performance.

Liv Langfeldt is research professor at the Nordic Institute for Studies in Innovation, Research and Education (NIFU), and director of the Centre for Research Quality and Policy Impact Studies (R-QUEST, www.r-quest.no). Her fields of competencies include research policy, research quality and peer review processes.

Paul Wouters is professor of scientometrics at Leiden University and dean of the faculty of Social and Behavioral Sciences. He was the former director of the Centre for Science and Technology Studies (CWTS).