Measuring impact in research evaluations

Lutz Bornmann

Division for Science and Innovation Studies
Administrative Headquarters of the Max Planck Society
Hofgartenstr. 8,
80539 Munich, Germany.
E-mail: bornmann@gv.mpg.de
Abstract

Governments all over the world are contemplating the question of where they should distribute public money (to education or to defence, for example). Distribution of money over a number of different areas always makes an issue, implicitly or explicitly, of the impact which can be achieved with investment in any one of them. Science is also affected by this governmental interest in impact; the issue is not only the impact of research on research itself, but on other areas of society. Citations are traditionally used to measure the impact of research on research. It is as yet unclear how the impact of research on other areas of society can be measured. It appears that alternative metrics (altmetrics, such as Twitter counts) might play a key role in this. This paper is concerned with the measurement of citation impact and societal impact, and looks at the basis, the effects and the problems of impact measurement.

Key words

Citation impact; Societal impact; Altmetrics
1 Introduction

Governments all over the world are contemplating the question of how they should distribute public money to different areas (such as building and maintaining infrastructure, educating children and young people and protecting the natural world nationally and internationally). Distribution of money over a number of different areas always makes an issue, implicitly or explicitly, of the impact which can be achieved with the investment in any one them (Morgan, 2014). Questions arise, such as will an investment in protecting the natural world create a better environment for humans and increase species diversity? Does investment in education have a positive impact on the strength of the national economy by bringing well-educated young people and adults into the labour market? As science also competes with other areas in society for public money, it is also faced with the challenge of demonstrating its value to society. Basic research in particular undergoes close scrutiny for this purpose: scientists can appreciate its value to society, but the general public (or politicians) can hardly do so (Bornmann, 2012, 2013).

The general public (or on its behalf, politicians) focuses on solutions to real-world issues (such as water and energy usage or the protection of the environment) and would like to know the extent to which research is contributing to the work done in these areas. Thus, they do not want to know only about the general impact of research on society, but also about the specific impact on socially relevant issues (Finkel, 2014; Thwaites, 2014). It could be said that with its interest in the impact of research on these real-world issues, society is regaining an understanding of the purpose of scientific work which is no longer an aspect of contemporary life. Modern society has separated into different sections (such as the economy, religion and justice) which are relatively autonomous, operate as separate entities and have their own laws (Luhmann, 2012a, 2012b). That research measures its own impact on research is in accordance with the notion of the autonomous nature of the different parts of society.
Measuring the impact of research in a way that relates to specific socially relevant issues is counter to this notion of autonomy and can mean focussing on subjects about which science would normally not be concerned (because there is nothing more to research in that area, for example).

In a number of countries (including Australia, Belgium, France, Italy, New Zealand, and the UK), national evaluation systems have become established which require science to account for funding and show that the money invested in it has not been wasted. These systems evaluate not only the research impact of research institutions on scientific progress, but also on the economy, environment, defence and public health. Since 2014 the UK, for example, has had the Research Excellence Framework (which replaced the Research Assessment Exercise, RAE) which evaluates the universities using peer review, case histories and metrics (Thwaites, 2014). Excellence in Research for Australia (ERA) is structured similarly to the REF in the UK. “ERA assessments are made by internationally recognized researchers, organized by discipline and clustered into eight Research Evaluation Committees. They use traditional measures of quality, such as citation analysis or peer review, but also incorporate a broader view, considering income from commercialization and measures of esteem – for example being admitted to a learned society such as becoming a fellow of the Australian Academy of Science” (Thwaites, 2014, p. S60).

The national evaluation systems are used, on the one hand, to evaluate national research in a global context. The main question is whether national research performs better or worse than the world average. The systems are used, on the other hand, to allocate research funds which are allocated more strictly in accordance with competition criteria than in the countries which do not have a system of this kind. The use of these systems is particularly interesting for small economies, where financial resources are more limited than in larger economies. In small economies, there is more pressure to use the little money available as profitably as possible.
2 Measurement of impact

Even though metrics and case studies are being used more frequently to evaluate research in various countries, the backbone of modern research evaluation has been and remains the peer review process in which colleagues mutually evaluate their research output (Bornmann, 2011b). Peer review is the oldest method of research evaluation and its use is closely associated with the development of modern science. It is only when research results are monitored by qualified experts in the same field to ensure that they meet certain standards that a research area can make reliable and valid statements. However, peer review comes up against its limits when a large number of research units need to be reviewed. In national evaluation systems, as a rule nearly every university and non-university research institution in a country undergoes scrutiny. To evaluate a larger number of units, quantitative methods have therefore been developed in addition to peer review, which use certain indicators to measure the output and/or the impact of research.

The most important quantitative method is currently bibliometrics, in which citations are used to measure the impact of publications. For example, the Australian Research Council (ARC) uses mainly citation analysis to evaluate research in the natural sciences. It is only in the engineering and social sciences and in the humanities that the ARC uses the peer review of selected publications or outputs, as citation analysis is inadequate for these disciplines. According to Sheil (2014), “it has demonstrated that citation analysis can be used as a principle indicator of quality in many disciplines, producing finely grained information about research strengths and weaknesses”. Bibliometrics is the favoured method of research evaluation primarily because papers and books are the most important product of science. It is only when research results are recorded in writing that they become accessible to a wider circle of recipients and can become the subject of critical discourse. According to Karl R. Popper (1961), only rational discourse conducted among fellow researchers can lead to
scientific progress. It is only when research results are published under the name of a researcher in a paper that the results are associated with this name and other researchers can refer to the author's achievements by using the results. If the citation impact of published research results is measured in science, a causal relationship is established between the two works: the citing work is influenced intellectually in one way or another by the cited work (in most of the cases, see Bornmann & Daniel, 2008). For example, citation counts at the level of the individual author provide information about how useful the research output of this author has been for science.

So far, no standard indicators have emerged that can measure the benefit of research to society (that is, the broader impact of research) reliably and with validity (Campbell & Grayson, 2014). The search is ongoing in scientometrics for a "citation equivalent", with which "activities such as working in industry, contributions to government reports or communication of research outcomes to audiences other than a researcher’s peers” (Thwaites, 2014, p. S59) can be measured. “Citation equivalents could be calculated for issued patents, commercial contracts and licence fees. More broadly, citation equivalents could be awarded for activities including writing books, opinion pieces and government submissions, PhD student supervision, and development of new approaches to teaching practices or novel training courses” (Finkel, 2014). Currently, the universities mainly use case studies to determine the value of their research for society (in the UK's REF, for example). However, the use of case studies in evaluation is criticised because it is believed that this approach can be used very selectively to show results, if the universities report only those results with the best impact (Morgan, 2014). Furthermore, case studies are expensive and time-consuming to prepare and assess and it is extremely difficult to generalise their results (Sheil, 2014).

The problem of finding a suitable method with which to measure societal impact also indicates that, all in all, it is much more difficult to measure societal impact than scientific impact. This difficulty arises primarily because there are many different target groups for the
impact on society (while scientific impact is always the impact on science). Societal impact can be the impact on policy makers, business, culture or other more specific sections of society. To measure societal impact, therefore, a decision needs to be made in each instance about what that impact actually is (Morgan, 2014). In the past, two methods of measuring societal impact based on a citation-equivalent have emerged as particularly useful: evaluating citations in patents (technological impact) and in clinical guidelines (medical impact). Both approaches have the following advantages: (1) societal impact can be measured in a similar way to scientific impact (and a sound method of evaluating the data – developed over decades of scientometrics research – is therefore available). (2) The fact that they are based on citations means that non-reactive, relatively objective and extensive data is available. (3) Patents and guidelines are available for the evaluation in a relatively freely accessible form and – compared with other data – can be evaluated with a reasonable amount of effort.

For example, Narin, Hamilton, and Olivastro (1997) studied the frequency with which scientific publications were cited in US patents. They evaluated 400,000 US patents issued between 1987 and 1994. Their results show that the knowledge flow from US science to US industry tripled in these years. Grant (1999) and Lewison and Sullivan (2008) pursued a similar objective to Narin, et al. (1997) with their evaluation of clinical guidelines: how does knowledge flow from clinical research to clinical practice? The pilot study by Grant (1999) examined three guidelines and was able to ascertain that they contained citations of a total of 284 publications (which can be categorised by author, research institution, country, etc.). For Grant (1999), the study results demonstrate the usefulness of his approach to tracing the flow of knowledge from research funding into clinical practice. Lewison and Sullivan (2008) substantially expanded the data basis studied by Grant (1999) and examined 43 UK guidelines for references to papers. The authors found the following: “The UK papers were cited nearly three times as frequently as would have been expected from their presence in world oncology research (6.5%). Within the United Kingdom, Edinburgh and Glasgow stood out for their
unexpectedly high contributions to the guidelines’ scientific base” (Lewison & Sullivan, 2008, p. 1944).

As well as clinical guidelines and patents, another source of data has emerged in recent years with which the impact of research in society could be measured: alternative metrics (altmetrics). “Altmetrics … is a term to describe web-based metrics for the impact of scholarly material, with an emphasis on social media outlets as sources of data” (Shema, Bar-Ilan, & Thelwall, 2014). In "article-level metrics" (ALMs), counts of views, downloads, clicks, notes, saves, tweets, shares, likes, recommends, tags, posts, trackbacks, discussions, bookmarks, and comments are counted, rather than just citations of a paper in a database such as Scopus (Elsevier), or by a publisher such as the Public Library of Science (PLOS) (Liu, Xu, Wu, Chen, & Guo, 2013; Zahedi, Costas, & Wouters, 2014). These ALMs are log data which measures individual mentions (such as a download) over a certain period of time (Haustein, 2014). “Today, for every single use of an electronic resource, the system can record which resource was used, who used it, where that person was, when it was used, what type of request was issued, what type of record it was, and from where the article was used” (Kurtz & Bollen, 2010, p. 4).

The "Almetric for Institutions" tool from the company Altmetric allows institutions to track, monitor and report on the broader impact of research (http://www.altmetric.com/institutions.php). It counts, analyses and, with some processing, presents online mentions of publications issued by an institution on platforms such as Mendeley, Twitter, CiteULike and Facebook. However, Altmetric analyses more than mentions on social media platforms. Its analyses also draw on government policy documents and other sources for mentions of scholarly articles. Particularly government policy documents are a important source with which to measure the societal impact of research as the impact of research on policy making can be thus made quantifiable (Liu, 2014). This systematic way of measuring impact on policymaking is potentially of interest in the social
sciences and the humanities in particular. In these disciplines, it is almost impossible to apply traditional bibliometrics and researchers in scientometrics are looking for alternative ways of evaluating research (Hug, Ochsner, & Daniel, 2013). As we can assume, for example, that social science studies which traditionally deal with issues such as unemployment, study and career paths and poverty in a specific country are hardly received on an international level, but are referred to on a national level by governments and other state bodies, measurement of impact using altmetrics in a specific area represents an interesting option for research evaluation in the social sciences.

Publications (and not software, devices, datasets, presentations etc.) will play a key part in the measurement of societal impact, as they do in the measurement of impact on science. Publications offer the best platform with which to convey research results to a specific audience. While traditional articles are directed at fellow researchers in the same discipline and therefore appear in the format with introduction, methods, results and discussion sections, publications directed at people outside of science should have a different format which is tailored to the target group in question. For example, results from Bornmann (in press) show that publications which are a key paper in a field, are well written, provide a good overview of a topic, and/or are well suited as literature for students accumulate higher altmetrics counts (particularly Twitter counts) than other papers. The results are an initial indication that a certain type of papers – irrespective of their quality – has an advantage in their reception beyond science.

Bornmann and Marx (2014) have taken this idea – that there should be a certain type of paper tailored to a readership outside of science to achieve societal impact – and proposed that scientists should write an 'assessment report' of such papers summarising the status of the research on a certain subject. These reports should represent knowledge available for society to access, and should be couched in generally understandable terms so that readers who are not familiar with the subject area or the scientific discipline can make sense of it. Even though
scientists in the discipline would seem to be the obvious choice to write these assessment reports, the task could equally be undertaken by others, such as journalists with the necessary specialist knowledge. To ensure that the reports are of high quality, they should undergo peer review to determine their correctness. The reviewers are asked to recommend the publication or rejection of the report and possibly with formulating suggestions for improvement to the submitted work (Bornmann, 2011b). Since the report will be read by scientists from other fields and non-scientists, it should be reviewed not only by an expert in the field but also by a stakeholder from government, industry or an advice centre. Societal impact of an assessment report is given when its content is addressed outside of science (in a government document, for example).

According to Thwaites (2014), health researchers are particularly interested in the impact that their research has beyond science. Health researchers would like to exert influence on medical practitioners with their research. Furthermore, government is very interested in the impact of research on health care as the cost of research in this area is very high. As measuring societal impact – not just in the area of health care – is (still) very difficult and the impact is often only felt after several years, proposals have been made in societal impact research to measure the endeavours (of institutions) to achieve impact and to honour them accordingly. Instead of "monitoring impact", it would be more of a matter of "monitoring progress towards impact" (Morgan, 2014). This change of perspective would mean, regarding assessment reports for example, that not the impact of assessment reports, but the number and quality of these reports would be measured. Many Dutch organisations involved in quality assurance cooperate in a project entitled Evaluating Research in Context (ERiC), which has set itself the goal of developing methods for societal impact assessment (ERiC, 2010). One significant result of the project is that productive interaction is a necessary requirement for research to have a societal impact: “There must be some interaction between a research group and societal stakeholders” (ERiC, 2010, p. 10). Such interactions can be in the form of
personal contact (e.g. joint projects or networks), publications (e.g. educational and assessment reports) and artefacts (e.g. exhibitions, software or websites). All these interactions can be counted as activities to achieve social impact.

3 Effects of impact measurement

Since bibliometric indicators have obtained a general acceptance in science policy and attained applied relevance in research evaluation, adaptation strategies by scientists resulting from the use of these indicators for science funding decisions have been reported (Evidence Ltd., 2007; Lawrence, 2003). Bornmann (2011a) has proposed designating these strategies "mimicry in science". Adaptation strategies can cast general doubt on the system of research evaluation: Goodhart’s law states that “once a measure becomes a target it ceases to be a good measure” (McGilvray, 2014, p. S66). There is a risk that scientists are more concerned about the marketing of their research products than the content of their research. Scientists apply strategies that enable them to comply with bibliometric accountability and to secure funds for their own research. Some of these strategies:

(1) Researchers do not summarise the results of a project in a single publication, but distribute the results over many publications to create the appearance of higher productivity (Bornmann & Daniel, 2007; Mallapaty, 2014). This publication strategy is referred to as "salami slicing". Results from Adams and Gurney (2014) for the UK Research Excellence Framework show that willingness to publish in books has steadily declined over recent years.

(2) As scientists as a rule strive to be judged as world class, they try to get their papers published in prestigious journals (Finkel, 2014). Therefore, they do not choose those journals which are most suitable for the subject matter, but those which are most highly regarded by their colleagues. In addition to these changes to publishing behaviour, there is also (3) an assumed pressure on scientists “to change research focus to better align with mainstream or more highly esteemed fields” (McGilvray, 2014, p. S64).

(4) Greater reluctance in established
researchers to work with early career researchers (as they often publish in less prestigious journals), and (5) a tendency in institutions to appoint academics who already have a long publication record instead of (younger) talented researchers who do not (McGilvray, 2014).

Looking at the literature on the effects of impact measurements, it is hard to find a discussion of the possible positive effects of measuring productivity and citation impact; this is not the case with the negative effects. As the positive effects should also be mentioned, there follows a description of some of them. Firstly, it is doubtless deemed a positive feature of a research unit if its publication output increases and does not decrease. Even though the increase of publications could be seen partly as a result of salami slicing, it does not obstruct the accumulation of knowledge in science. The form in which the results are published is of no significance to this accumulation; what is important is that they are published and can be accessed by fellow researchers. The orientation of scientists towards the possible impact of publications can also be considered in a positive light. Work on research topics is always work carried out by a relevant community to which research groups from various institutions and countries make their own contribution. One way of determining in how far a research group has made a useful contribution is whether their publications are cited by other research groups or not. The research group should therefore aspire to be cited by other groups, as these citations reflect the usefulness of their research. To increase the visibility of their own research results, it is important for researchers to publish in journals that are as prestigious as possible (within the discipline).

Establishing measurement of the societal impact as part of research evaluation – alongside measuring the impact on science – would lead to scientists’ angling their work more towards groups of people outside of science. For example, the assessment reports described above would direct descriptions of research results at a specific group of readers outside of science. With alternative metrics it would then be possible – as with traditional citations – to measure the impact of the reports outside of science. If there were a desire to implement a
system of output measurement (number of assessment reports) and impact (number of mentions of the reports) like this in a national evaluation system, it would be necessary to consider in advance whether this kind of measurement would have the intended effects. (1) Is the expected adaptation in scientists' behaviour, for example that they write reports for other status groups in society in addition to producing their research papers, a desirable one? (2) Is science making a greater contribution to solving real-world issues (such as energy usage or environmental protection) with these new products, and do the economy, the environment, defence and public health in a country (see above) gain anything from it? Do the new products result in a benefit for science which would not be expected without the products? According to Morgan (2014), the development of a suitable evaluation methodology, which takes account of these and similar questions is a “a big sticking point” (p. 75) which scientometrics should address.

David Sweeney (Director for Research, Innovation and Skills at the Higher Education Funding Council for England, HEFCE, in Bristol), who managed development of the REF, pointed out at a symposium in February 2014 in Melbourne, Australia, (on issues surrounding the outcomes and impact of how research is measured) that research evaluation can also result in the acquisition of other, i.e. private sources of money (Thwaites, 2014). Under matched-funding schemes, the results of research evaluations can be used to supplement public funding for research with private money. The successful acquisition of private money requires that the success of research is documented both inside and outside science. It is therefore advantageous for a research institution to use impact figures to demonstrate its particular strengths not only to public funders (which is indispensable in evaluated research) but also to other potential providers of finance.
4 Problems of impact measurement

Evaluations cost time and money, irrespective of how they are structured. A science system with evaluation must be significantly more successful than a science system without evaluation in order to be able to compensate for the disadvantage in terms of time and money. To keep time and money costs low in evaluations, it is without a doubt more helpful to give more weight to quantitative over qualitative methods as part of an informed peer review process. This is the method used by REF in the UK and by Excellence in Research for Australia, ERA. As the measurement of societal impact has arisen as a new (quantitative) element in evaluations over recent years, the question arises of how important it really is to include the measurement of societal impact in a national evaluation system, in terms of the balance of cost of and yield from evaluations.

Aidan Byrne, Chief Executive at the Australian Research Council (ARC), reported at the Melbourne symposium mentioned above that the results of the ERA programme for academic excellence correlate with other real-world outputs. “For instance, 95% of industry investment in research in Australia is in the same areas in which researchers performed at world-class or better. And the same is true for 98% of the research that was commercialized and for 97% of the work that was patented” (Thwaites, 2014, p. S60). If there really is a correlation between the economy in a country and excellent research, then one could claim that measuring the impact of research on society is actually obsolete: excellent research has a societal impact anyway (on the economy) – it is merely necessary to create the conditions in a country under which research can be carried out at the highest level.

Science is a part of society marked by inequality, random chance, anomalies, the right to make mistakes, unpredictability and a high significance of extreme events. As this section is intended to illustrate, these characteristics can lead to problems with measuring the impact of scientific performance.
Inequality: The instance of inequality is very high in science. In almost every set of papers (by a researcher or an institution) there are many papers which are not cited at all or are hardly cited and only a few highly cited papers (Seglen, 1992). As the findings of Bornmann, de Moya-Anegón, and Leydesdorff (2010) illustrate, papers contributing to scientific progress in a discipline lean on to a larger extent the few previously important contributions than on papers contributing little. However, it is not only on the level of individual papers, but also on the level of individual researchers that we see great inequality affecting publication output and citation impact. Ioannidis, Boyack, and Klavans (2014) investigated 15,153,100 publishing scientists (distinct author identifiers) in the Scopus database. Their results show that “only 150,608 (<1%) of them have published something in each and every year in this 16-year period (uninterrupted, continuous presence [UCP] in the literature) This small core of scientists with UCP are far more cited than others, and they account for 41.7% of all papers in the same period and 87.1% of all papers with >1000 citations in the same period”.

The right to make mistakes and the significance of extreme events: Science is characterised not only by unequal distribution, but also by the right to make mistakes. According to Popper (1961), scientific issues, hypotheses and problems are solved through trial and error (and not with the empirical confirmation of previously formulated hypotheses). When a problem is examined by researchers, the attempts at solution are tested empirically, whereby the poor alternatives are identified and removed. Scientific research is therefore always prone to error and associated with the right to use fallible trial and error, to take risks and unrecognised (unorthodox or intuitive) routes. Scientific progress based on trial and error is not however as a rule cumulative; i.e. the ongoing production and elimination of knowledge units, but happens through extreme events which Kuhn (1962) has called scientific revolutions. Dramatic publications in a discipline lead to completely new thinking, which
relates primarily to the taxonomy used in an area (Wray, 2011). The taxonomy used before a revolution is fundamentally different from the taxonomy after the revolution.

**Anomalies**: Where inequality, error and extreme events are important components in a system, we can assume that anomalies also play a key role. As the studies by Bornmann, et al. (2010) and Ioannidis, et al. (2014) cited above have shown, science is determined essentially by a few elements (such as publications and scientists) and not by a mass of many. Even the journals which are continuously analysed for the Web of Science (Thomson Reuters) are selected on the assumption that scientific progress can be represented by a few core journals and therefore the vast majority of journals need not be included. As bibliometric analyses are as a rule undertaken on a higher aggregate level (such as institutions or countries), the effect of anomalies (the few highly cited papers) on the whole is underestimated or they are treated as problems. For example, Göttingen University only achieved a good place in the current Leiden ranking (which uses a mean-based indicator to rank institutions) because it could publish one extremely highly cited publication in recent years (Waltman et al., 2012). With analyses on a higher aggregate level, where data is very skewed, a few anomalies are responsible for the results but their immense influence on the overall result (such as the average citation impact of an institution) is frequently not visible. An institution does not acquire an excellent score in an impact measurement on the basis of the arithmetic average of citations due to the majority of scientists it employs and their publications, but due to its few very successful scientists or their small number of highly cited papers.

**Randomness and unpredictability**: Campanario (1996) has published some good evidence for the existence of serendipity in science: He "re-examines 205 Citation Classics commentaries from the 400 most-cited papers in the recent history of science. Authors of 17 Citation Classics commentaries (8.3%) mention some kind of serendipity in performing the research reported in the highly cited paper.” The term 'serendipity' is closely associated in scientific research with the name of Robert K. Merton who published on this topic with Elinor
Barber (Merton & Barber, 2004). Much that is of importance in science is the result of happy coincidence – arising from unknown causes or causes which are almost impossible to find rationally. The discovery of X-rays and penicillin are good examples of this. The existence (and importance) of random elements in scientific work indicate that there are obviously parallels between the acquisition of knowledge in science and evolutionary processes in nature. While living beings adapt better with random genetic changes in nature, scientific findings are the result (among other things) of studies which are by chance appropriate. Researchers find things that they were not even looking for.

It follows from this randomness in the process of finding knowledge that important progress in science is often almost completely unpredictable. There are many examples in science where the importance of certain research results has appeared only decades after their publication (van Raan, 2004). At the time of publication, no one was expecting that it would have any significance for the scientific community. For example, the “Shockley-Queisser limit” (Shockley & Queisser, 1961) describes the limited efficiency of solar cells on the basis of absorption and reemission processes (Marx, 2014). The original reception of the paper in terms of citations was initially hesitant. Until 2014, however, the paper has become one of relatively few highly cited papers in its field. The citations of the paper published by Shockley and Queisser (1961) developed relatively synchronously with the rapidly growing solar cell and photovoltaic research area.

However, even when results have been presented in a discipline and only later proved to be particularly important, experts in the same field frequently do not recognise their wider significance and therefore do not cite these papers. Marx and Bornmann (2010, 2013) showed in two bibliometric studies on scientific revolutions that a number of publications, which turned out to be important for a revolution, were rarely cited. An important discovery is therefore often credited to the author who links the various strands of knowledge required for the discovery in a meaningful way and not those who have contributed or published a crucial
empirical result or theory. Bornmann and Marx (2012) have proposed the designation "Anna Karenina Principle" for this process where different and necessary strands of knowledge derived from theoretical and empirical processes are brought together to create a revolutionary discovery: "Success in science at the excellence level always depends on several key aspects that must all be fulfilled. One missing aspect can lead to failure (science at most at the medium level) and makes the failure a unique matter” (p. 2047).

The importance of inequality, anomalies, randomness, unpredictability, error and revolutionary events in the process of scientific discovery is frequently seen in literature as evidence that science cannot be planned, managed or measured and therefore it would be superfluous to measure impact within the framework of research evaluations. Even if the cost of evaluations is high, there would not be commensurate benefit for the process of scientific discovery. Ziman (2000) writes, for example: “It is a matter of some concern that post-academic researchers have much less freedom to undertake relatively modest investigations without formal approval. In effect, they are all deemed to be doing ‘normal’ science, which has no place for serendipity. Research grants and contracts tie them into finalized projects from which they are not officially permitted to deviate to chase after presumed wild geese” (p. 218). Similarly Morgan (2014) writes: “Many scientific discoveries are serendipitous, and critics suggest that … a goal-oriented approach [which characterizes evaluation frameworks] could lead to less blue-sky research funding” (p. 72).

The question of in how far impact measurements damage or benefit the process of scientific discovery (or have no effect on research performance) remains unanswered. Although research evaluations are conducted at great cost, meta-evaluations to test the effectiveness of evaluations are hardly ever undertaken. One of the few examples of a meta-evaluation is the study by Bornmann, Mittag, and Daniel (2006), in which the processes evaluating studying and teaching in Germany is meta-evaluated. The fundamental question to be asked here is undoubtedly whether a science system which is barely affected by research
evaluations is marked more strongly by important indicators such as anomalies, chance, error and revolution than a science system which is subject to more evaluation. On the one hand, it is easier to envisage these characteristics in a deregulated science system. The more accurately the research process is planned and measured, the less opportunity there seems to be for randomness, error and revolution. However, it is also conceivable that there is higher probability of randomness, error and revolution in an evaluated science system which has an extreme focus on excellence. Planning and carrying out research can be two sides of the same coin (here excellent research) with different intentions.

In addition to the characteristics of science (such as randomness, error and revolutionary events in the process of scientific discovery), which can lead to problems in the measurement of impact as part of research evaluations, the following discusses four other aspects which can be problematic for impact measurement. These four aspects are (1) the elements usually evaluated in a measurement of impact; (2) the continuity desired in research performance; (3) the lack of interval estimation and (4) evaluating research within disciplines.

**Evaluated elements:** Even when research is carried out by individuals or smaller groups, the impact measured is frequently that of institutions (such as universities) or countries (such as the BRICS states). University rankings such as the Academic Ranking of World Universities (http://www.shanghairanking.com), which report regularly on the performance of research institutions, are a good example (Hazelkorn, 2011). As the results of Bornmann, Mutz, and Daniel (2013) based on the universities considered in the Leiden Ranking show, only 4.3% of the variance in institutional citation impact can be attributed to differences among the universities. The rest, namely 95.7%, is allocated to variance within the universities (that is, to departments, research groups, and individual researchers). The heterogeneity of research performance is therefore so pronounced within the universities that it seems questionable to measure impact at institutional level. Instead, it seems more appropriate to evaluate research performance on the basis of smaller units. In this way, it
might be possible to ascribe the research achievement to those who were originally responsible for it.

A possible reason for the popularity of impact analyses at institutional and country level is undoubtedly the availability of data. It is much harder to collect data on research groups and individual scientists than on institutions and countries. The data on individuals is difficult to identify because different people share the same names. While authors of papers always give the name of the institution where they work and the country where the institution is located, they rarely provide any information about the research group, the department or similar. Another reason for the popularity of institutions and country analyses is undoubtedly the political focus on these units. The government of a country sees itself in competition with other countries and therefore calls for the relevant country-based studies. Furthermore, the basic funding of research is generally provided via the university and not via an individual scientist or research groups. Governments therefore like to be informed how the money has been used to benefit the whole university.

**Continuity**: Generally speaking, the evaluation of science is continuous. Institutional evaluations are repeated every few years. Each new evaluation of an institution is expected to show that the output and impact of the previous evaluation has been at least achieved, and preferably surpassed. Scientists at an institution are therefore asked to produce a constant and increasing flow of important research results which can be published in high-impact journals and later achieve a high impact. This expectation does not correlate with the (current) working process in science. It is rather a long and laborious period of generating ideas, carrying out studies and (statistical) analysis of data, the visualisation and recording of results and discussion of them with colleagues. Each of these stages is expensive and mistakes and setbacks can occur. Even though there are hardly any statistics on the subject, we can nevertheless assume that many scientific studies make slow progress, and are not completed because they are too arduous.
Despite this situation, in evaluations it is nevertheless expected that this work is designed as a continuous process which generates a regular flow of output and impact. It is expected that a steady input of financial resources will create steady output and impact. However, it is not usually possible to assume this linear relationship between input and output in scientific work. If the relationship is assumed to be linear, there is the danger that the disparity between the aspiration of the evaluation and the reality of the research process results in scientific misconduct to adjust the reality to the aspiration. To counter the dissatisfaction at this disparity, misconduct is taken into account for, or in, an evaluation.

Merton's (1938) general theory on aberrant behaviour in society can be used to explain the connection between dissatisfaction and aberrant behaviour in science. Merton (1938) distinguishes between three factors in social structure to explain aberrant behaviour in society (such as criminality in the USA): (1) Certain wishes and expectations represent important cultural goals in a society; (2) norms which state rules for how these goals should be reached and (3) the distribution of resources required to reach the goals. For Merton (1938), aberrant behaviour (particularly among the 'underprivileged') results when a social structure makes it impossible to achieve the cultural goals shared and internalised by a society (in the USA, this would be individual wealth, for example) with socially accepted means (such as honest work). Aberrant behaviour is likely to occur in persons in a certain group if they can only achieve the goals dictated by society with great difficulty if they use legitimate means. It arises in a situation in which certain symbols for success are strongly overemphasised by the group, but only a small part of these symbols can be acquired by legitimate means, due to the way that resources are distributed.

Against the background of this general mechanism to explain aberrant behaviour, scientific misconduct would result from a situation in which the researchers are faced with the goals set up by the institution ('winning the game') which can only be reached with great difficulty, or not at all, within the cognitive and social norms in science ('winning through
circumscribed modes of activity'). The research results are falsified or massaged in order to produce for continuous evaluation new and sensational results which are published in highly regarded journals and later highly cited. McGilvray (2014) describes an attempt at gaming in a national evaluation system: “Perhaps the most egregious example of an attempt to game the system occurred in New Zealand in 2006. One leading university reclassified dozens of staff members, notably those who were PBRF [Performance-Based Research Fund] - eligible but performed little active research. By reclassifying inactive researchers away from subjects such as economics and biology to fields such as philosophy and religious studies, the university would improve its standing in the former fields. The surge in the number of New Zealand philosophers piqued the curiosity of PBRF reviewers who eventually reversed the classifications” (p. 66).

The recourse to aberrant behaviour (such as fabricating and falsifying research results and reassigning people for an evaluation) can be attributed to (1) excessive emphasis on objectives in the continuous evaluations, which are defined by certain symbols of success in research (such as higher citation impact); (2) a generally diminished importance of rules intended to be applied to the process of achieving the objectives (such as achieving a high citation impact) and (3) the restricted availability of financial and personnel resources with which to achieve the objectives.

**Error rate:** The Journal Impact Factor (JIF) is one of the most important indicators in bibliometrics and is used both for measuring the performance of journals (for example, by libraries deciding on the purchase of journals) and for the evaluation of individual papers or researchers. There is criticism of the use of the JIF, not only because it is deemed unsuitable for measuring the impact of individual papers and researchers, but also because it is calculated with such precision – to three places after the decimal point – which cannot be achieved with the analysis of bibliometric data. With the analysis of impact data (to calculate the JIF) there is a number of sources of error in the data: (1) the information supplied by the authors on their...
publications is incorrect; (2) not all the publications are included in the databases; (3) the bibliographic information in the publications is transferred incorrectly to the databases; (4) the data is read out from the database with errors and (5) errors are made in the analysis of the data. The criticism expressed about the overly precise representation of the JIF can also be applied to other impact indicators (Fischhoff & Davis, 2014). Bibliometric indicators are frequently presented as precise figures without any information about possible sources of error.

For bibliometrics, one can argue that the observed citation impact of papers allows making inferences about the underlying process that generated those impacts and the extent to which citations may have been influenced by random factors. The citation success of a paper is presumably affected by the quality of the results in the papers, but is also partly determined by chance. How often a paper or collection of papers gets cited might be affected both by the errors mentioned above and also by how many people chose to read a particular issue of a journal, who happened to learn about a paper because somebody casually mentioned it to them (Bornmann & Daniel, 2008; Williams & Bornmann, 2014). Therefore, if the same paper were to be published at another time in the same journal, one could expect similar but not identical results for the citation impact.

Generally speaking, if we could somehow repeat the publication of the paper and its citation analysis over and over again, the citation impact would not be exactly the same for each repetition, just like doing 100 coin tosses over and over would not yield the exact same number of heads each time. To demonstrate the random nature of bibliometric results it is important to present the bibliometric results with an interval estimate (Williams & Bornmann, in press). Up to now it has been usual to present the result as a single value (a point estimate) determined from the bibliometric data. However, since we can assume that the point estimates (such as an average citation impact) fluctuates from random sample to random sample, it is desirable to make statements about the accuracy of the estimate. Confidence intervals, for
example, are useful here, to give a range in which one can very confidently expect the actually unknown value (such as the average citation impact of an institution) to be.

Providing confidence intervals with bibliometric results and their use in the interpretation of the results means that the point estimate presented (such as e.g. the average citation impact of those papers which an institution has published in a certain period) is not viewed as a precise figure. For example, in university rankings, such as the Leiden Ranking (www.leidenranking.com), universities are ranked by bibliometric impact scores. This can result in only a marginal difference in the performance scores for the universities; however, these marginal differences can result in big differences in the ranking positions. As the users of the rankings (such as the general public or science policy-makers) are primarily interested in the ranking positions of universities, they are frequently unaware that the performances of the universities are actually similar, due to the potential errors described above. The use of confidence intervals and also stability intervals, as used by the Leiden Ranking (Waltman, et al., 2012), are measures to highlight the uncertainty of the results and the similarity of performance and to allow the results to be interpreted appropriately.

**Disciplinary research evaluation:** In connection with the benefits of research, McGilvray (2014) draws attention to a problem with research evaluation today which also affects impact measurement: "Researchers undertaking interdisciplinary work are … feeling compromised. Campuses across New Zealand and Australia are bringing together researchers from multiple disciplines, from the hard sciences to the humanities, to look at societal problems in a holistic way. These fields include environmental sustainability and medical research and, in many cases, work is carried out under the auspices of a centre or an institute within a university. Despite this big-picture approach, assessments such as the PBRF and the ERA continue to view research through a mono-disciplinary lens. The final report from the 2012 PBRF conceded that the 42 subject areas under which all research is assessed ‘do not accurately reflect the way research activity is organized and conducted’” (McGilvray, 2014, p.
S65). Using two examples, the following highlights in how far the mono-disciplinary lens can lead to problems in impact measurement.

(1) Nowadays, advanced bibliometric indicators are used when impact is measured with citations. They compare the impact of each paper with a reference value (Vinkler, 2010). Usually, the papers used for reference values for a paper are those which are in the same subject category (the same journal set) and have been published in the same publication year. However, this approach can be problematic with papers which have been generated by an interdisciplinary project. There are no appropriate subject categories for these papers. Using the normal subject categories can result in the publications from the project being compared with unsuitable reference sets, which can lead to an over- or underestimate of the impact of publications from interdisciplinary projects. For the publications from these projects, one would have to use topic-specific and not field-specific reference values (such as all the publications on climate change) in which the impact of publications on a certain topic is compared.

(2) Thomson Reuters recently published at highlycited.com the names of researchers who have published the most highly cited papers in their discipline (Bornmann & Bauer, in press; Thomson Reuters, 2014). It follows that the more papers a researcher publishes in a subject category (such as social sciences or agricultural sciences) which are in its 1% most-cited papers, the higher their chance of being included in the Thomson Reuters list. However, scientists whose work is interdisciplinary and who publish in very different subject categories are at a disadvantage compared to those working in a single discipline when it comes to being included in the list. As their highly cited papers are counted in different subject categories, they can be among the highly cited researchers in these different categories; however, they have less chance of prevailing over the mono-disciplinary researchers in the individual categories. Since the data are used for the Academic Ranking of World Universities, the
disadvantage of inter-disciplinary working scientists has also consequences for the position of universities in this important university ranking.

5 Discussion

There is no doubt that there would be no progress in society without science. Scientific research results in new technology, a greater understanding of our planet and the universe and in treatments which allow us to live longer and more healthily (Campbell & Grayson, 2014). Scientists have successfully investigated the reasons for climate change and worked on more secure and sustainable solutions to provide us with energy. They have significantly improved the accuracy of weather forecasts and made a substantial contribution to the control of infectious diseases. Despite these achievements, today's audit society (Power, 1999) expects that science and other areas of society are accountable to the state for their outcomes and that they address the specific problems which governments believe confront society with greater dedication. As these specific demands on science represent a relatively new phenomenon (arising in recent decades), there is a question of how they are changing or have changed science. In the view of Luhmann (2012a, 2012b), we can expect that science will take on these challenges and apply itself more to them; however, this will only be possible with the methods, instruments and practices with which it has always investigated the world. Research as an activity would not, therefore, in the view of Luhmann (2012a, 2012b) change fundamentally.

 Even if the way in which new findings are made probably does not change, it is likely that the way they are published will. If the impact of research not only on science but more widely across various areas of society is measured, it would be advantageous to formulate the findings for the relevant readership. Appropriately formulated results will have a better chance of creating an impact than texts which have not been worked on (that is, papers published in scientific journals). This preparation could be undertaken by others, such as
science journalists. However, journalists are at home in their own sub-section of society, the mass media (Luhmann, 2000), and operate according to the rules which apply there in order to reach a wider audience. If this preparation were to be undertaken by the scientists themselves, their publishing habits would have to change. In this context, Bornmann and Marx (2014) have proposed the assessment reports described above, in which researchers sum up the findings in their area so that they can be understood by a non-scientific readership. The writing of these reports can be designated as countable attempts of productive interactions, whereby the societal impact of the reports could be measured using altmetrics.

While scientometricians some years ago had to worry only about developing metrics with which to measure the impact of research on research, today the focus is on societal impact, which means measuring impact in very different sections of society. This research is essential, as there is already a number of countries – Australia, New Zealand and the UK – which use a metrics-oriented evaluation system with a broad measurement of impact. We can expect that more countries will choose these systems, as a metrics-based system is cheaper and simpler to run than a system based primarily on estimates by experts (Mallapaty, 2014; McGilvray, 2014) and furthermore, the systems appear to result in science achieving a better performance. According to Campbell and Grayson (2014), "both Australia … and New Zealand … have seen their global scientific standings rise in recent years – attributable at least in part to their assessment systems, even though Australia’s system offers little financial reward." The results from Bornmann and Leydesdorff (2013) indicate that the UK has improved the citation impact of its publications substantially over recent years and has now achieved a standard better than that of the USA. However, there are not as yet any comprehensive studies which could offer firm evidence of a link between an evaluation system and the attainment of more knowledge.

This paper has examined the measurement of impact and investigated the reasons why the impact of science is measured in a certain way. It has also considered the problems of
impact measurement and its unintended effects. Both areas should be accorded particular attention nowadays as impact measurement is (or should be) undertaken on a wider scale than it used to be. Also, the strong focus on excellence which has been observable in recent years in research evaluation can be problematic particularly for early career researchers, as Peter Gluckman, Chief Science Advisor to the Prime Minister of New Zealand, illustrated with this example: “I’ve seen several young researchers quite compromised by this drive to produce the one paper that will get into Nature,” says Gluckman, “when their career would have been much more developed had they focused on getting solid, excellent papers in the appropriate journals” (McGilvray, 2014, p. S65).

In-depth scientometric research is required to define metrics-based evaluation systems sensibly, effectively and systematically; however, this research should not only look at the development of reliable and valid indicators, but also at the effects of these systems. A key question should be in how far these systems can really result in science performing better.
References

Adams, J., & Gurney, K. A. (2014). *Evidence for excellence: has the signal overtaken the substance? An analysis of journal articles submitted to RAE2008*. London, UK: Digital Science.

Bornmann, L. (2011a). Mimicry in science? *Scientometrics, 86*(1), 173-177. doi: 10.1007/s11192-010-0222-8.

Bornmann, L. (2011b). 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*(8), 673-676.

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*(2), 217-233.

Bornmann, L. (in press). Validity of altmetrics data for measuring societal impact: A study using data from Altmetric and F1000Prime. *Journal of Informetrics*.

Bornmann, L., & Bauer, J. (in press). Which of the world's institutions employ the most highly cited researchers? An analysis of the data from highlycited.com. *Journal of the Association for Information Science and Technology*.

Bornmann, L., & Daniel, H.-D. (2007). Multiple publication on a single research study: does it pay? The influence of number of research articles on total citation counts in biomedicine. *Journal of the American Society for Information Science and Technology, 58*(8), 1100-1107.

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

Bornmann, L., de Moya-Anegón, F., & Leydesdorff, L. (2010). Do scientific advancements lean on the shoulders of giants? A bibliometric investigation of the Ortega hypothesis. *PLoS ONE, 5*(10), e11344.

Bornmann, L., & Leydesdorff, L. (2013). Macro-indicators of citation impacts of six prolific countries: InCites data and the statistical significance of trends. *PloS one, 8*(2), e56768. doi: 10.1371/journal.pone.0056768.

Bornmann, L., Marx, W. (2012). The Anna Karenina principle: a way of thinking about success in science. *Journal of the American Society for Information Science and Technology, 63*(10), 2037-2051. doi: 10.1002/asi.22661.

Bornmann, L., & Marx, W. (2014). How should the societal impact of research be generated and measured? A proposal for a simple and practicable approach to allow interdisciplinary comparisons. *Scientometrics, 98*(1), 211-219.

Bornmann, L., Mittag, S., & Daniel, H.-D. (2006). Quality assurance in higher education – meta-evaluation of multi-stage evaluation procedures in Germany. *Higher Education, 52*(4), 687-709.

Bornmann, L., Mutz, R., & Daniel, H.-D. (2013). A multilevel-statistical reformulation of citation-based university rankings: the Leiden Ranking 2011/2012. *Journal of the American Society for Information Science and Technology, 64*(8), 1649-1658.

Campanario, J. M. (1996). Using citation classics to study the incidence of serendipity in scientific discovery. *Scientometrics, 37*(1), 3-24. doi: 10.1007/bf02093482.

Campbell, P., & Grayson, M. (2014). Assessing science. *Nature, 511*(7510), S49-S49. doi: 10.1038/511S49a.

ERiC. (2010). *Evaluating the societal relevance of academic research: a guide*. 


Evidence Ltd. (2007). *The use of bibliometrics to measure research quality in UK higher education institutions*. London, UK: Universities UK.

Finkel, A. (2014). Perspective: Powering up citations. *Nature, 511*(7510), S77-S77. doi: 10.1038/s11192-012-0741-6.

Fischhoff, B., & Davis, A. L. (2014). Communicating scientific uncertainty. *Proceedings of the National Academy of Sciences, 111*(Supplement 4), 13664-13671.

Grant, J. (1999). Evaluating the outcomes of biomedical research on healthcare. *Research Evaluation, 8*(1), 33-38.

Haustein, S. (2014). Readership metrics. In B. Cronin & C. R. Sugimoto (Eds.), *Beyond bibliometrics: harnessing multi-dimensional indicators of performance* (pp. 327-344). Cambridge, MA, USA: MIT Press.

Hazelkorn, E. (2011). *Rankings and the reshaping of higher education. The battle for world-class excellence*. New York, NY, USA: Palgrave Macmillan.

Hug, S. E., Ochsner, M., & Daniel, H.-D. (2013). Criteria for Assessing Research Quality in the Humanities – A Delphi Study among Scholars of English Literature, German Literature and Art History. *Research Evaluation, 22*(5), 369-383.

Ioannidis, J. P. A., Boyack, K. W., & Klavans, R. (2014). Estimates of the Continuously Publishing Core in the Scientific Workforce. *PLoS ONE, 9*(7), e101698. doi: 10.1371/journal.pone.0101698.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago, IL, USA: University of Chicago Press.

Kurtz, M. J., & Bollen, J. (2010). Usage Bibliometrics. *Annual Review of Information Science and Technology, 44*, 3-64.

Lawrence, P. A. (2003). The politics of publication. Authors, reviewers and editors must act to protect the quality of research. *Nature, 422*(6929), 259-261.

Lewison, G., & Sullivan, R. (2008). The impact of cancer research: how publications influence UK cancer clinical guidelines. *British Journal of Cancer, 98*(12), 1944-1950.

Liu, C. L., Xu, Y. Q., Wu, H., Chen, S. S., & Guo, J. J. (2013). Correlation and Interaction Visualization of Altmetric Indicators Extracted From Scholarly Social Network Activities: Dimensions and Structure. *Journal of Medical Internet Research, 15*(11), 17. doi: 10.2196/jmir.2707.

Liu, J. (2014). New Source Alert: Policy Documents. Retrieved September 10, 2014, from http://www.altmetric.com/blog/new-source-alert-policy-documents/

Luhmann, N. (2000). *The Reality of the Mass Media*. Stanford, CA, USA: Stanford University Press.

Luhmann, N. (2012a). *Theory of Society* (Vol. 2). Stanford, CA, USA: Stanford University Press.

Luhmann, N. (2012b). *Theory of Society* (Vol. 1). Stanford, CA, USA: Stanford University Press.

Mallapaty, S. (2014). Q&A Jane Harding: Individual approach. *Nature, 511*(7510), S82-S83. doi: 10.1038/s11192-012-0741-6.

Marx, W. (2014). The Shockley-Queisser paper – A notable example of a scientific sleeping beauty. *Annalen der Physik, 526*(5-6), A41-A45. doi: 10.1002/andp.201400806.

Marx, W., & Bornmann, L. (2010). How accurately does Thomas Kuhn’s model of paradigm change describe the transition from a static to a dynamic universe in cosmology? A historical reconstruction and citation analysis. *Scientometrics 84*(2), 441-464.

Marx, W., & Bornmann, L. (2013). The emergence of plate tectonics and the Kuhnian model of paradigm shift: a bibliometric case study based on the Anna Karenina principle. *Scientometrics, 94*(2), 595-614. doi: DOI 10.1007/s11192-012-0741-6.
McGilvray, A. (2014). Research assessment: The limits of excellence. *Nature, 511*(7510), S64-S66. doi: 10.1038/511S64a.

Merton, R. K. (1938). Social structure and anomie. *American Sociological Review, 3*(5), 672-682.

Merton, R. K., & Barber, E. G. (2004). The travels and adventures of serendipity: a study in historical semantics and the sociology of science. Princeton, N. J. USA: Princeton University Press.

Morgan, B. (2014). Research impact: Income for outcome. *Nature, 511*(7510), S72-S75. doi: 10.1038/511S72a.

Narin, F., Hamilton, K. S., & Olivastro, D. (1997). The increasing linkage between US technology and public science. *Research Policy, 26*(3), 317-330. doi: 10.1016/s0048-7333(97)00013-9.

Popper, K. R. (1961). *The logic of scientific discovery* (2nd ed.). New York, NY, USA: Basic Books.

Power, M. (1999). *The Audit Society: Rituals of Verification*. Oxford: Oxford University Press.

Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science, 43*(9), 628-638.

Sheil, M. (2014). Perspective: On the verge of a new ERA. *Nature, 511*(7510), S67-S67. doi: 10.1038/511S67a.

Shema, H., Bar-Ilan, J., & Thelwall, M. (2014). Do blog citations correlate with a higher number of future citations? Research blogs as a potential source for alternative metrics. *Journal of the Association for Information Science and Technology, 65*(5), 1018-1027. doi: 10.1002/asi.23037.

Shockley, W., & Queisser, H. J. (1961). Detailed Balance Limit of Efficiency of P-N Junction Solar Cells. *Journal of Applied Physics, 32*(3), 510-&. doi: Doi 10.1063/1.1736034.

Thomson Reuters. (2014). The world's most influential scientific minds. Bethesda, MD, USA: Thomson Reuters.

Thwaites, T. (2014). Research metrics: Calling science to account. *Nature, 511*(7510), S57-S60. doi: 10.1038/511S57a.

van Raan, A. F. J. (2004). Sleeping Beauties in science. *Scientoometrics, 59*(3), 467-472.

Vinkler, P. (2010). *The evaluation of research by scientometric indicators*. Oxford, UK: Chandos Publishing.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijsse, R. J. W., van Eck, N. J., . . . Wouters, P. (2012). The Leiden Ranking 2011/2012: data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology, 63*(12), 2419-2432.

Williams, R., & Bornmann, L. (2014). Sampling issues in bibliometric analysis. Retrieved January 31, from http://arxiv.org/abs/1401.2254

Williams, R., & Bornmann, L. (in press). The substantive and practical significance of citation impact differences between institutions: guidelines for the analysis of percentiles using effect sizes and confidence intervals. In Y. Ding, R. Rousseau & D. Wolfram (Eds.), *Measuring scholarly impact: methods and practice*.

Wray, K. B. (2011). *Kuhn's evolutionary social epistemology*. Cambridge, UK: Cambridge University Press.

Zahedi, Z., Costas, R., & Wouters, P. (2014). How well developed are altmetrics? A cross-disiplinary analysis of the presence of ‘alternative metrics’ in scientific publications. *Scientoometrics, 1*-23. doi: 10.1007/s11192-014-1264-0.

Ziman, J. (2000). *Real science. What it is, and what it means*. Cambridge, UK: Cambridge University Press.