# Finding Research Value – the metrics and methods for evaluating research

Michael J. Freeman

## Introduction – How can we evaluate the value of research?

The value of the research done by individuals is used by institutions as part of the process of allocating resources to that individual; those resources being salary, job title, tenure, use of facilities, etc. Similarly, the aggregate value of all the research coming from a research institution is sometimes used by funding institutions as part of the process of allocating resources.

Traditionally, the process of evaluating the value of research has largely depended on bibliometric methods, such as citation counts or journal impact factors. However, there are many criticisms of the accuracy, or even the appropriateness, of these various bibliometrics, when used for valuing research (cf. MacRoberts and MacRoberts 1989; Kostoff 1998; van Raan *et al.* 2007; Butler 2008; Adler *et al.* 2009; van Eck *et al.* 2013). Perhaps more importantly, bibliometric methods do not capture the wider value of research—the "full range of economic, social, public policy, welfare, cultural and quality-of-life benefits" (Grant *et al.* 2009) that can result from research (van Raan *et al.* 2007). A common theme in the literature is that "the future of research assessment endeavors lies in the intelligent combination of metrics (including bibliometric indicators) and peer review" (Moed 2009).

In the following, first I detail some of the most popular bibliometric methods, how they work, and the criticisms of their use in evaluating the "research impact"—as opposed to the "socio-economic impact" or "practical impact"—of an individual researcher, research group or research institution.

I then describe some examples of how organizations and institutions have tried to, or have at least recommended as a way to, combine the evaluation of research impact and practical impact in order to arrive at an understanding of the "complete impact of research". It should be noted that the primary intent for some of those evaluation frameworks (e.g., the Research Quality Framework in Australia) was to provide input on the allocation of funds to research institutions (Donovan 2008). If the influence of practical impact on the allocation of resources to research institutions increases, presumably its influence on the allocation of resources to individual researchers will also increase.

## Bibliometrics

The bibliometric methods used as a measure of—or at least as a surrogate for the measure of—the impact of research can be divided into two general classes: journal-level metrics, and article-level metrics.

Journal-level metrics rank the journals in a field of research. It is generally assumed that any article published in a journal with a high "impact-factor" is of high quality, and therefore researchers with publications in high impact-factor journals are doing high quality research.

Article-level metrics are more direct, and aggregate the data about all of an individual's published articles into a single measure, in order to give an indication of the quality of that researcher's work.

## Journal-level metrics

Journal-level metrics are the earliest metrics to become widely used, and continue to be influential. They measure the impact of journals in a research field using citation counts between journals. Journal-level metrics were originally developed to aid librarians in selecting which journals to which to subscribe and to maintain in their collections.

But journal-level metrics have become used as a way to indicate the quality of a researchers publications. Presumably, the logic goes, a journal must publish the best articles in a field in order to have a high impact factor, so therefore any article published in a journal with a high impact factor must be among the best articles.

### Journal Impact Factor

All citation-based metrics can probably be traced back to a paper by Gross and Gross (1927); Archambault and Larivière (2009) tap that article as the starting point for the Journal Impact Factor. Since the Journal Impact Factor is the best known and possibly most widely used citation-based metric for research, its origins can serve as an example for all journal-level metrics.

**Finding the "indispensable" journals** The question posed by Gross and Gross (1927) was, "What… scientific periodicals are needed in a college library?" They wanted to identify the journals to which a library should subscribe, and they used quantitative methods to make comparisons between journals, and those methods eventually led to the Journal Impact Factor.

Gross and Gross decided to use a quantitive method to evaluate journals, as opposed to merely asking an expert to compile a list of "indispensable" journals in order to avoid a list that was "seasoned too much by the needs, likes and dislikes of the compiler." In other words, they used quantitative methods to minimize any single individual's biases from the evaluation of journals.

**Counting citations to sort journals** The basic method pioneered by Gross and Gross, and repeated later by other researchers, was to select a pre-eminent journal or a reference that was generally accepted as key to the field at hand, and then compile and quantify the sources cited in that keystone journal/reference.

This is an extremely simplified example: if the journal that was selected as the central reference for a field, "Field", contained five citations to a "Journal A", ten citations to a "Journal B", and four citations to a "Journal C", then for that field the journals would be ranked and reported something like this:

Table 1: Leading periodicals in Field

| Journal | Citation count |
|---|---|
| Journal B | 10 |
| Journal A | 5 |
| Journal C | 4 |

These early citation metrics were inherently field-specific, since they took data from a central reference to a particular field.

**Compiling citations for multiple fields** Soon, though researchers began to cross-compile citation information from multiple journals, and then to include journals from multiple fields in those compilations. For example, Gregory (1937) compiled more than 27,000 citations from across 27 subfields in medicine.

Gregory gives a concise explanation of the purpose of her metrics:

> The foregoing Tables (1-27) answer primarily the needs of the specialist in his attempt to keep adequately abreast of the literature in his field. Two further Tables have been compiled, to indicate to the medical library and the librarian (A), the indispensable periodicals for all the fields consulted; (B), a short list of essential periodicals in general medicine which cover a large amount of material. This latter list is designed primarily for the individual and for the small library.

There were 27 "foregoing Tables", one for each of the subfields, yet in Gregory's Herculean compilation, there were no comparisons made between fields. This is a feature that would continue in similar metrics for decades: when multiple fields were compiled and evaluated at the same time, the results for each field were reported separately. Presumably this was because the intended audience for these metrics had no need for cross-field comparisons, since libraries are generally tasked to select the best journals for some particular field, and are not given carte blanche to choose which fields to support, and researchers are interested in their field, not all fields.

The compilation of citations from Martyn and Gilchrist (1968) is cited by Archambault and Larivière (2009) as having a large influence on methods used in the Journal Impact Factor, and it too continues the practice of reporting the metrics for separate fields separately.

However, when journal impact factors eventually become used as a way to judge the quality of research, this lack of cross-field comparisons becomes an issue as institutions try to equitably evaluate the quality of research done in an array of fields. Ways to address this problem are offered by some journal-level metrics as refinements to the methodology of the Journal Impact Factor. Some of these are described in later sections.

**Reporting ratios instead of counts** Another feature of the Martyn and Gilchrist (1968) compilation that was carried forward into the Journal Impact Factor and other bibliometrics was first proposed by Hackh (1936): reporting a ratio instead of a count. Specifically, the ratio of the number of citations to the number of pieces that might be cited.

Consider another extremely simplified example: if there were five citations to articles published in "Journal A" in the same year that "Journal A" published a total of 20 articles, the citation ratio would be "5:20", or "0.25".

We can extend our earlier simplified example to include reporting ratios instead counts, and see why this is the preferred method:

Table 2: Citation ratios – simplified example

| Journal | Citation count | Citable pieces | Ratio |
|---|---|---|---|
| Journal B | 10 | 20 | 0.50 |
| Journal A | 5 | 20 | 0.25 |
| Journal C | 4 | 10 | 0.40 |

Table 3: Leading periodicals in Field

| Journal | Ratio |
|---|---|
| Journal B | 0.50 |
| Journal C | 0.40 |
| Journal A | 0.25 |

Despite having fewer citations than Journal A, Journal C has a higher ratio of citations to citable pieces, and so is ranked higher. Using a ratio to rank journals is a way to prevent quantity from overwhelming quality.

**And thus the Journal Impact Factor** This, then, is the core of how the Journal Impact Factor and a number of similar metrics work. Gather citation counts and citable pieces for each journal, calculate the ratios, and rank the journals.

However, there are a number of possible problems with this method. For example, the Journal Impact Factor has been criticized for the following reasons (among others):

*The Journal Impact Factor…*

*…is not sensitive to differences in composition of the journals.* Review pieces in journals that aggregate recent research into an overview nearly always have higher citation rates than research articles. Letters and notes tend to have higher citation rates in the short-term. Differing journal impact factors may therefore represent nothing more than a difference in number of each type of citable documents in the journals (van Leeuwen *et al.* 1999).

*…uses too short of a citation window.* For subfields of mathematics, social sciences and humanities "the use of an indicator which depends on a citation window of one to two years seems to be clearly insufficient to measure the impact of research in those areas" (van Leeuwen *et al.* 1999).

*…is is not sensitive to differences in fields.* The Journal Impact Factor does not account for differences in "the size of professional communities, the numbers of their indexed journals and type of articles in different fields, and the researchers' differing citing behaviours" and therefore, impact factors cannot be appropriately compared between fields. (Bornmann *et al.* 2012)

So while the Journal Impact Factor remains popular, other journal-level metrics have been developed that either address specific problems or implement newer methodologies, e.g., prestige rankings. Three examples of those are described in the following section.

## Alternatives to the Journal Impact Factor

### SNIP – Source Normalized Indicator of journal impact per Paper

As mentioned earlier, the lack of cross-field comparisons in the Journal Impact Factor means that its use is problematic for institutions that want to equitably evaluate the quality of research done by faculty and staff in an array of fields. The Journal Impact Factor of journals cannot be compared between fields "because citation practices can vary significantly from one field to another" (Moed 2011).

The source normalized indicator of journal impact per paper, 'SNIP', includes refinements to address this problem through the normalization of citations. SNIP normalizes the citation rates of articles against the average number of cited references per paper in that subject field (Moed 2010). In a simplistic example, if the articles in a field typically cite 10 articles, then a set of articles with a non-normalized citations rate of 12 would have a normalized citation rate of 1.2, which can be characterized as above average for that field.

To normalize away the differences in citation practices, however, requires properly delineating where those differences exist; in an article responding to criticisms of SNIP, Moed (2010) cites Garfield (1979) who reported that citation practices differ not only between fields, but between specialties and sub-specialties.

To ensure that a journal's citations are normalized against an appropriate field of research, within SNIP a particular journal's field is not determined by categorization but instead it is defined by what articles cite the journal. In other words, the "field" is the collection of all articles *outside* the journal that contain citations to articles *inside* the journal. (It should be noted that the term "article" is being used loosely

here. A more accurate description might be "citable document", since within SNIP "*articles*, *conference proceedings papers* and *reviews* are considered as fully fledged, peer-reviewed research articles" (Moed 2010).) In short, SNIP normalizes a journal's citations based on the field's average citations, where what constitutes the field is defined by the citations to the journal.

A version of SNIP that runs against the Scopus® database is available on the Journal Metrics website, *www.journalmetrics.com* (Elsevier 2015).

### SJR

Another problem with the Journal Impact Factor is that all citations count the same, regardless of their source. However, some would argue that citations from certain sources should be weighted more than others, much like the opinion of an expert carrying more weight than that of a novice. The SCImago Journal Rank indicator (SJR), does just that by using an implementation of social network analysis to rank the relative prestige of the journals. This more gradated ranking is useful because more "poorly cited journals are entering the indices, [therefore] it is essential to have metrics that will allow one to distinguish with greater precision the level of prestige attained by each publication" (González-Pereira *et al.* 2010).

The SJR calculates for each journal its eigenvector centrality, which is "a measure of centrality… in which a unit's centrality is its summed connections to others, weighted by their centralities", where 'centrality' is "network-derived importance" (Bonacich 1987) . For the SJR, therefore, 'centrality' is an indicator of a journal's prestige. The eigenvector centrality calculated for a journal is then normalized using the total number of citations to the journal, resulting in a size-independent rank.

The resulting SJR metric is "aimed at measuring the current 'average prestige per paper' of journals for use in research evaluation processes" (González-Pereira *et al.* 2010).

### PageRank / Eigenfactor.org

Another way to rank journals based on prestige instead of simple popularity, is the approach used by Eigenfactor.org®, which "ranks the influence of journals much as Google's PageRank algorithm ranks the influence of web pages" (West 2015).

Bollen *et al.* (2006) examined using the PageRank methodology to rank journals by prestige, using "the dataset of the 2003 ISI Journal Citation Reports to compare the ISI [Impact Factor] and Weighted PageRank rankings of journals." To achieve the ranking, citations are used as a basis to "iteratively" pass prestige from one journal to the other, until "a stable solution is reached which reflects the relative prestige of journals." This is an adaptation of the original PageRank method, the difference being that connections between journals are weighted, based on citation frequencies.

Prestige and citation rankings were found to be largely similar, with two notable exceptions. The first were "journals that are cited frequently by journals with little prestige" that rank much lower on a prestige index than they do on a citation index. The second were the converse of the first, "journals that are not frequently cited, but their citations come from highly prestigious journals" so that they rank much higher on a prestige index than they do on a citation index.

Based on specific examples of these two types of journals that have distinctly different rankings when sorted by citations versus prestige, in general it is theory-heavy journals that have low citation counts, yet high prestige. The journals with high counts but low prestige are "methodological and applied journals" or ones "that frequently publish data tables" (Bollen *et al.* 2006).

It is, however, unclear if this result is an indication of the importance of including work on theory to produce high-quality research, or simply evidence that a weighted PageRank methodology is effective at maintaining previous perceptions of the relative values of theoretical and applied research.

## Journal-level metrics are poor surrogates for research evaluation

The variety of variants to the Journal Impact Factor might at first seem to be an indication that there is a lot of competition to win the favor of librarians. Because, of course, journal-level impact metrics like the Journal Impact Factor were born of the desire to rank the importance of journals so that librarians will know which ones to have in their libraries.

And while modern journal-impact bibliometrics may be distantly related to the original work by Gross and Gross (1927), they have the same focus: ranking the journals themselves. The leap seems to be huge, from journal rankings to judgements of the quality of a individual articles in the journals, but in practice, it's often unnoticed. For example, in the introduction to an article describing yet another variant on the Journal Impact Factor, the authors write:

> The citedness of a scientific agent has for decades been regarded as an indicator of its scientific impact, and used to position it relative to other agents in the web of scholarly communications. In particular, various metrics based on citation counts have been developed to evaluate the impact of scholarly journals…. (González-Pereira *et al.* 2010)

In other words, since the number of citations received by a researcher, or research group, can be a useful indicator of the quality of their research, it is useful to count the number of citations received by the ~~researchers~~ journals in which the researchers publish. The logic seems to be reversed: instead of extrapolating that good articles make the journals they are in better, we extrapolate that good journals somehow make the articles that are in them better.

As Adler *et al.* put it:

> …Instead of relying on the actual count of citations to compare individual papers, people frequently substitute the impact factor of the journals in which the papers appear. They believe that higher impact factors must mean higher citation counts. But this is often not the case! This is a pervasive misuse of statistics that needs to be challenged whenever and wherever it occurs. (Adler *et al.* 2009)

Yet, it's hard to explain the variety of bibliometrics available to assess the impact of journals, and the nature of some of the refinements that differentiate them, without assuming that the metrics are being used for something more generally desirable than the ranking of the journals themselves.

The problems inherent in using journal-level metrics as an indicator of article quality can be avoided by counting the citations of the article themselves. This leads us to the other class of bibliometrics used to evaluate the quality of research, article-level metrics, which are discussed in the next section.

## Article-level metrics

An obvious way to overcome the logical and practical problems of using journal-level metrics to evaluate the quality of the articles, is to base evaluations on the articles themselves. Various article-level metrics are available to do this.

### *h*-Index

> Here, I would like to propose a single number, the '$h$ index,' as a particularly simple and useful way to characterize the scientific output of a researcher. (Hirsch 2005)

The "$h$-index" is indeed a simple metric. To find the $h$-index of a researcher, take all the articles that the researcher has published, and sort them in ascending order of how many citations each has received. Then start counting the articles, starting with the one with the greatest number of citations; when you come to an article which has a number of citations equal to the count of articles so far, that article/citation count is the $h$-index for the researcher. For example, a researcher who has an $h$-index of 6 has published six articles with at least six citations each.

The $h$-index is popular enough to have prompted the creation of various refinements to the basic method, e.g., the g-index (Egghe 2013) and the $h^{\mathrm{m}}$-index (Schreiber 2008), and to define two of the three citation metrics displayed on Google Scholar's user-citations page. Those three metrics are: the user's $h$-index, the i10-index, a variant of the $h$-index (Connor 2011), and finally a count of total citations.

Yet, the usefulness of the $h$-index is unclear. In the original proposal, the index was offered as a quantitative metric to be used "for evaluation and comparison purposes" (Hirsch 2005). To demonstrate this use the author reported the $h$-index value for a collection of example researchers, including Nobel-prize winners. But, as Adler *et al.* explain:

> One can conclude that it is likely a scientist has a high h-index given the scientist is a Nobel Laureate. But without further information, we know very little about the likelihood someone will become a Nobel Laureate or a member of the National Academy, given that they have a high h-index. (Adler *et al.* 2009)

Indeed, the $h$-index has be shown to be inferior to equally simple metrics: "Compared with the h-index, the mean number of citations per paper is a superior indicator of scientific quality, in terms of both accuracy and precision." (Lehmann 2006).

Though, there is a more important question to ask than which of the simple, article-level metrics is a better indicator of research quality—are any of them actually useful? Evaluating the quality of the work being done by a researcher is not a simple task, so any simple tool will inevitably prove to be inadequate. In reference to the h-*index* and a number variants of it, Adler *et al.* write:

> These are often breathtakingly naïve attempts to capture a complex citation record with a single number. Indeed, the primary advantage of these new indices… is that the indices discard almost all the detail of citation records, and this makes it possible to rank any two scientists. Even simple examples, however, show that the discarded information is needed to understand a research record. (Adler *et al.* 2009)

Despite the complexity of citation records, simple, single-number metrics comprise the majority of commonly referenced article-level metrics. This mirrors the wide array of single-number journal-level metrics available. Perhaps this focus on lone-value quantification is a consequence of having electronic indices of research publications that make such quantification simple.

Presumably, using data sources not so readily reduced to a single number would produce a more complex measure of a researcher's work. Such measures are classed as "alternative metrics," a.k.a. "altmetrics", and are described next.

### Altmetrics

Though technically not a bibliometric method, altmetrics have been used to perform analyses similar to the bibliometric methods already discussed, so they are included here.

Altmetrics use data available on the web such as "usage data analysis (download and view counts); web citation, and link analyses" (Zahedi *et al.* 2014) are used to supplement and improve upon citation-based metrics for measuring the impact of science and research. Using information made available via the web,

altmetrics can go beyond the journal articles and books included in citation-based metrics and include "other outputs such as datasets, software, slides, blog posts, etc." (Zahedi *et al.* 2014).

So, instead of compiling citations to research in journals, altmetrics involves compiling "mentions" of research in "main-stream media sources, and social media shares and discussions", along with statistics on downloads, and reference manager counts (Adie and Roe 2013).

Altmetrics also typically retain the related meta-data of mentions and usage statistics, which allow for more complex analyses of the information. In other words, altmetrics not only track what research is being "mentioned", but also where it is mentioned, and who is mentioning it, which can potentially provide a richer understanding of the citations.

Perhaps most importantly, by drawing on sources outside of academia and research journals, altmetrics might provide some indications of the "pratical impact" of research. This is discussed in the "Future research" section.

**Bibliometrics, scientometrics, and informetrics** If altmetrics aren't bibliometrics, what are they? They are "informetrics" which is a broader field than either bibliometrics, or "scientometrics". Bibliometrics has been defined as "the quantitative study of physical published units, or of bibliographic units, or of surrogates of either" (Broadus 1987). Scientometrics has been defined as "the study of the quantitative aspects of science as a discipline or economic activity", which includes the practice of publication and citation, so it "overlaps bibliometrics to some extent" (Tague-Sutcliffe 1992). Informetrics, in contrast, includes quantitative studies of not only quantitative methods applied to publications, but also documentation and information (Egghe and Rousseau 1990); it has also been described as a "recent extension of the traditional bibliometric analyses also to cover non-scholarly communities in which information is produced, communicated, and used" (Ingwersen and Christensen 1997).

For a more detailed discussion of the differences between the three areas, see "The Literature of Bibliometrics, Scientometrics, and Informetrics" by Hood and Wilson (2001).

## Conceptual issues with citation analysis

It's important to remember the origins of the citation-based journal-level metrics, which are also the conceptual ancestors of citation-based article-level metrics. Specifically, what's important is that the citation-based metrics were developed to aid librarians in deciding what journals would be most useful to their patrons (Gross and Gross 1927, Archambault and Larivière 2009). And in research libraries the patrons are likely to be among the authors of the articles in those journals; in other words, the patrons are likely to be the people who make citations in the first place. So, prioritizing journals by citation rates makes sense, since the journals most likely to contain citable documents in the future are those that contained citable documents in the past.

The transition from using citation-based metrics for selecting what journals to carry in a library to using those metrics for evaluating the "research impact" of articles was based on the assumption that the most important research will necessarily get cited the most. Yet this is obviously not always true. For example, the article "Electrochemically induced nuclear fusion of deuterium" (Fleischmann and Pons 1989) received 490 citations in its first four years after publication, but 'cold fusion' (as it came to be know) had little actual research impact (Moed 2005).

So, citation-based metrics do not account for the nature of citations—a negative citation counts the same as a positive one. To further complicate matters, the ability to clearly distinguish a negative from a positive citation, and the ability to state that a negative citation doesn't represent a net positive influence in the field, are dependent on how you define "intellectual influence". Indeed, this dependence extends to citation-based metrics in general:

> The citation impact of [a] work can be interpreted in terms of intellectual influence. If one disregards the permanence of the intellectual influence, its cognitive direction and its longer term implications, this concept becomes more similar to that of citation impact. On the other hand, if an evaluator considers these aspects of intellectual influence as important attributes in a qualitative assessment, discrepancies between a work's citation impact and the assessment of its intellectual influence are apt to rise. (Moed 2005)

This helps explain why, in the examples of the evaluation of research impact in later sections, expert panels shied away from using citation-based metrics as anything more than an 'indicator' of research impact. "In order to be useful and properly used in research evaluation, citation impact must be further interpreted" within a framework that is appropriate for the field, and the particular research in question (Moed 2005). This is what was proposed for the Australian Research Quality Framework, and what was implemented in the United Kingdom's Research Excellence Framework, which are described in the sections on practical impact that follow.

# Indicators of practical impact – examples from government initiatives and programs

As mentioned in the description of altmetrics above, the biggest limitation of bibliometrics, when used to judge the value of research, is that they are limited to the impact that research has on other research. To get to the total impact of research, information from outside the realm of research must be included.

This is why there is a parallel shift in the nature of source materials when shifting topics from considering research quality to considering the practical impact of research. The outcomes of interest are no longer restricted to within the world of research, and so neither are the source materials.

Similarly, the examples used to discuss the evaluation of the practical impact of research are practical examples. They come from programs initiated and/or implemented by government agencies in order to make funding for research institutions be responsive, to some extent, to the practical impact of the research being produced by those institutions.

## Background on examples from government initiatives and programs

Both Australia and the United Kingdom have developed government programs that evaluate research institutions based on their "research quality" and "research impact". Those programs used "research quality" to refer to what is sometimes called the "academic impact" of research; it is a measure of the effect that research has on other research, and included some form of bibliometrics.

"Research impact" is the effect that research has *outside* of research and research institutions; it is an evaluation of the "full range of economic, social, public policy, welfare, cultural and quality-of-life benefits" (Grant *et al.* 2009) that can result from research. In this thesis, it is referred to as "practical impact".

### Research Quality Framework (RQF)

The history of the Research Quality Framework (RQF) is complicated. It started in 2003 when the Australian government established "an Expert Advisory Group, whose remit was to consult widely and develop a model for assessing the quality and impact of research in Australia" (Butler 2008). The framework proposed by that group was published in 2006, and prompted the establishment of the Development Advisory Group, to "refine the RQF model" (Donovan 2008). The first RFQ was scheduled to be run in 2008, but a change in government in 2007 instead meant that the program was scrapped (Donovan 2008).

However, the Research Quality Framework has influenced the development of similar programs, such as the Research Excellence Framework in the UK (Grant *et al.* 2009), and discussion of its development are available in the literature (cf. Butler 2008, Donovan 2008).

#### Research Excellence Framework (REF)

The Higher Education Council for England "funds and regulates universities and colleges in England" (HEFCE 2015).

The HEFCE will be using the Research Excellence Framework (REF) "for the assessment and funding of research in UK higher education institutions". The REF focuses on three elements (HEFCE 2009a):

1. **Research quality** – research "will be assessed against criteria of 'rigour, originality and significance'. By 'significance', we mean the extent to which research outputs display the capacity to make a difference either through intellectual influence within the academic sphere, or through actual or potential use beyond the academic sphere, or both."
2. **Research impact** – "…demonstrable economic and social impacts that have been achieved through activity within the [research institution] that builds on excellent research."
3. **Research environment** – "…the extent to which the [research institution] has developed a research infrastructure, and a range of supporting activity, conducive to a continuing flow of excellent research and to its effective dissemination and application."

However, the three elements are not equally weighted: "The assessment of research impact will be one of three distinct elements of the REF, being judged alongside research excellence and research environment, contributing 25% towards the overall outcome (as compared with 60% and 15% for quality and environment)." (Technopolis 2010)

A pilot exercise of the REF was conducted in 2009, reports on the outcomes and evaluations from that pilot are referenced in the following sections.

## Evaluation of 'research quality' in the examples

#### REF

Much like with 'practical impact' (discussed below), the HEFCE ran a pilot exercise of the evaluation of research quality at higher-education institutions. Specifically aiming to validate the use of bibliometric methods, the key finding of the exercise was: "Bibliometrics are not sufficiently robust at this stage to be used formulaically or to replace expert review in the REF" (HEFCE 2009b). But it was not a complete rejection of bibliometrics, as "there is considerable scope for citation information to be used to inform expert review" (HEFCE 2009b).

Thus, the REF uses expert panels, comprising both experts in the discipline as well as end-users of research, to evaluate the research quality of an institution, with bibliometric methods being included in the evaluation process. (Grant *et al.* 2010)

This approach also aligns with the view given in a report commissioned by HECFE before the running of the pilot exercise from the Centre for Science and Technology Studies, Leiden University:

> Citation counts can be seen as manifestations of intellectual influence, but the concepts of citation impact and intellectual influence do not necessarily coincide. Citation impact is a quantitative concept that can be operationalised in an elementary fashion or in more sophisticated ways, such as crude citation counts versus field-normalised measures. Concepts such as 'intellectual influence' are essentially theoretical concepts of a qualitative nature, and have to

> be assessed by taking into account the cognitive contents of the work under evaluation. Thus, the outcomes of bibliometric analysis must be valued in a qualitative, evaluative framework that takes into account the contents of the work. (van Raan *et al.* 2007)

Using bibliometrics within a qualitative framework—such as review by expert panels—was also the approach in the final version of the RFQ.

**RFQ**

Under the Australian RFQ, research quality would have been assessed by a panel of experts, with one quarter of the panel members being end-users of research. The panel would have used a spectrum of bibliometric measures to compliment the review process.

The government ministry created a 'Quality Metrics Working Group', which developed recommendations for what bibliometrics would—and would not—be appropriate. The working group specifically rejected the Journal Impact Factor (né ISI Impact Factor):

> It was believed that actual citation counts are a far better citation measure for judging the performance of groups than surrogates based on the average citation rates of the journals which carry that work. There were also concerns about the way in which the indicator is calculated and anecdotal evidence of increasing manipulation of the indicator by a few journal editors. Even when ranking journals, some disciplines had already made it clear that they wished to look beyond the Impact Factor and undertake a more detailed assessment of the quality of journals. (Butler 2008)

The working group on metrics recommended that panels choose from a "suite" of metrics that included citations reported as simple counts, averages, or centile distributions. Also, it was recommended that some fields might want to include citations from "non-standard venues", meaning from outside the standard indicies of research publications from which bibliometrics typically draw data—though this practice was discouraged, since drawing citations from venues outside those indices would be a labor-intensive process. And the working group specifically recommended "that no attempt be made to aggregate the indicators to produce a single score" (Butler 2008).

Ostensibly the working group recommended against aggregating quantitative measures to alleviate concerns that a single quantitative measure might have undue influence, but the advice also falls in line with concerns that "analysis of research performance on the basis of journals unavoidably introduces a 'bibliometrically limited view of a complex reality' " (van Raan *et al.* 2007).

So, while bibliometrics were to play a role in the RFQ, and do play a role in REF, they are used within a qualitative framework since the evaluation of research is a complex and muli-faceted problem. As Adler *et al.* write:

> We do not dismiss citation statistics as a tool for assessing the quality of research—citation data and statistics can provide some valuable information. We recognize that assessment must be practical, and for this reason easily derived citation statistics almost surely will be part of the process. But citation data provide only a limited and incomplete view of research quality, and the statistics derived from citation data are sometimes poorly understood and misused. Research is too important to measure its value with only a single coarse tool. (Adler *et al.* 2009)

So, a mix of qualitative and quantitative methods is used to evaluate research quality. And it turns out that a mixture of methods is also used for the evaluation the practical impact of research, a discussion of which is in the next section.

## Evaluation of 'practical impact' in the examples

### REF

The documentation published by HEFCE for the REF explained that practical impacts would include "a wide definition of impacts, including economic, social, public policy, cultural and quality of life" and that any reference to " 'impact' or 'social and economic impact' " implicitly included the entire wide range of impacts. (HEFCE 2009a)

This did not prevent confusion about what was meant, unfortunately. In a "lessons learned" report about the HEFCE-REF pilot exercise, one of the items given as a challenge was explaining to academic groups "that socio-economic impact was a much broader concept than economic impact" (Technopolis 2010).

Some outside the process seem similarly unclear on the wider impact being sought by the REF. Commenting on the high ratings received by all groups taking part in the pilot exercise, Khazragui and Hudson (2015) write, "But in part too it is a consequence of having economic impact evaluated by non-economists." The implication being that since economic impact is the dominant feature of overall impact, an over-estimation of economic impact would inevitably cause a large-magnitude effect on the results. Therefore the high ratings can be explained by a failure to properly quantify the economic impact.

The objection of Khazragui and Hudson to a lack of quantitative methods, presumably is what led them to denigrate the nature of the evaluation process, writing, "research funders also illustrate their impact with 'stories' " (Khazragui and Hudson 2015).

The authors are presumably are referring to the "narrative evidence" that was used for the REF, since "there are limitations in the extent to which the impacts of research can be 'measured' through quantifiable indicators." The REF used a qualitative process.

> Rather than seek to **measure** the impacts in a quantifiable way, impact will be **assessed** in the REF. Expert panels will review narrative evidence supported by appropriate indicators, and produce graded impact sub-profiles for each submission; they will not seek to quantify the impacts. (HEFCE 2009a)

Quantitive data were included in the process, however. The HEFCE directed that the case studies that were submitted by research institutions should "include a range of **indicators of impact** as supporting evidence". Those indicators were expected to be quantified values, such as the research income generated from other funding sources, and accountings of collaborations with companies in the private sector (HEFCE 2009a).

### RFQ

In the RFQ, research would have practical impact if it created benefits in at least one of four domains: social, economic, environmental, or cultural (Donovan 2008)

- Social Benefit – new approach to social issues, improved policies, improved equity, improved health, safety and security, etc.
- Economic Benefit – improved productivity, increased employment, increased innovation and global competitiveness, "unquantified economic returns resulting from social and public policy adjustments", etc.
- Environmental Benefit – reduced waste and pollution, reduced consumption of fossil fuels, improved management of natural resources, etc.
- Cultural Benefit – greater understanding of "where we have come from, and who and what we are as a nation and society", stimulating creativity within the community, contributing to cultural preservation and enrichment, etc.

However, the definition of impact in the final version of the Australian RFQ program was the result of tensions between defining impact in terms of "the interests of industry and commerce" and defining it in relation to "broader public benefits" (Donovan 2008). These tensions are played out in the differences between three sources: two sets of preliminary recommendations for the program and the final version of the RFQ.

For example, one of the preliminary recommendations defined impact in terms of "direct practical utility" and within the document the term 'impact' was at times used interchangeably with the word 'usefulness'. It advocates measuring impact with quantitative metrics, such as business investments and returns, numbers of patents issued, etc. In contrast, the other preliminary recommendation promotes combining quantitative indicators with qualitative evidence in order to include the intangible benefits of research in the evaluation process. The final version of the RFQ included a "case study" methodology for evaluating impact, which would allow for the more holistic method of evaluation, but also included a scale for the scoring of impact that used magnitudes of impact as a criteria, which clearly favors quantitative evidence (Donovan 2008).

Despite the shift of focus between quantitative, economic benefits, and qualitative, social benefits, the RFQ never rejected the idea that practical impact was an important part of judging the overall value of research.

The REF and RFQ each combined qualitative and quantitative methods in order to evaluate both the quality of research and its practical impact, in order to assess the overall value of research.

# Future research

While there is an established history of evaluating the quality of research, often by using bibliometric methods, there is little precedent for evaluating its practical impact. The next two sections discuss possibilities in regards to that evaluation.

## Can altmetrics provide practical impact data?

Most bibliometric methods for measuring research quality only use data from indicies of academic journals, and use time windows of five years or less. They are therefore unlikely to capture the practical impact of research, because monitoring only academic publications means they can't detect practical impact directly, and the relatively short time windows means they can't detect the indirect influence that an important practical impact will eventually have on research.

Altmetrics can use data from a wider range of contexts, and so it should have the potential to capture some of the practical impact of research. Unfortunately, altmetrics don't exclude data from academia, either, which may obscure evidence of practical impact by diluting it with a large volume of data due to meticulous citation in altmetric "mentions" by researchers.

An example of both this potential and possible dilution is the study by Mohammadi *et al.* (2014) of Mendeley readership data. Mendeley is a reference manager that tracks what research is being read by its users, and what makes that data potentially useful for identifying practical impact is the inclusion of the users' "profession" in the meta-data. Users self-select their profession from a pre-populated list of options, so it's not infallible, but it does allow for classing users as being research academics (those who author citable research), non-research academics (those who don't author papers—e.g., students), or someone outside of academia. The research preferred by the class of users who are not members of academia would presumably be the research that is most relevant to practitioners, and the research that is having a practical impact.

As Mohammadi *et al.* put it: "It seems that Mendeley readership is able to provide evidence of using research articles in contexts other than for their science contribution, at least for Social Science and some applied sub-disciplines. …It also could be used as a supplementary indicator to measure the impact of

some technological or medical papers in applied contexts…" (Mohammadi *et al.* 2014). As both of the 'contexts' mentioned are forms of practical impact, what the author is describing is evidence of research having a practical impact.

It is important to note some of the caveats given by the authors. "Mendeley is perhaps most useful for those who will eventually cite an article and so its readership counts seem likely to under-represent users who will never need to cite an article, perhaps including disproportionately many practitioners" (Mohammadi *et al.* 2014). So, while the data from Mendeley suggests the potential of tracking practical impact, it may not be able to fulfill it. Also, "although the Mendeley API provides information related to the discipline, academic status and country of readers for each record, it only reports percentages rather than raw data and only gives information about the top three categories." So, if one of the top three categories of readers isn't a practitioner category, almost no information about an article's impact on practice ends up being available.

A similar potential for finding evidence of practical impact exists in social media sources as in Mendeley readership data. The various examples of Altmetric LLC data given by Adie and Roe (2013) hint at this potential for extracting practical impact information, while also highlighting some of the difficulties in utilizing those sources. E.g., though Adie and Roe were able to collect meta-data about the users involved in "mentions" of citable research, there's no indication that there is any data available which would allow users to be classified as being, or not being, members of academia.

But, then again, this lack of meta-data suitable for classifying users is only a hinderance to the automated analysis and quantification of data. Finding and flagging altmetric mentions in social-media discussion can provide leads to information about the practical impact of research. Which is presumably why Almetric LLC provides "data sources that can be manually audited by our users. If Altmetric [LLC] says that an article has been tweeted about five times, then users should be able to get the relevant five links, Twitter usernames, and timestamps to go check for themselves" (Adie and Roe 2013).

If institutions step away from using metrics and move toward using indicators in the evaluation of research quality and impact, the field of altmetrics would seem to be ripe with potential indicators to fit the bill.

## How can researchers be "ahead of the curve" in evaluations of research impact?

Even when not officially required, demonstrations of the practical impact of research can be a useful addition in any context when there is an evaluation of the value of research. Documenting or demonstrating impact can also be internally helpful, as "in doing so a great majority will derive insight and local benefits" (Technopolis 2010).

The Technopolis report on feedback from the HEFCE-REF pilot includes a lot that would seem to recommend that researchers develop a portfolio of research impact *before* it's required of them. Participants in the pilot exercise reported that, "It did cause departments to take a fresh look at their work, and it shed new light on the breadth of good things that have happened in part as a result of the excellent research being carried out. It has helped people to reflect on their wider contributions…". The irony in the report is that "The exercise has also revealed how little we do know about the use that is made of our research and equally that where one has a sense of good things happening, we simply don't record that anywhere. Our evidence base is pretty scant". This irony is echoed a number of times: research institutions were unaware of the wider impact they were having despite that impact being real and positive. E.g., "we were pleasantly surprised at the outcome of the case studies. These clearly provided a much broader appreciation of the impact the university's research has had / is having than previously recognised" (Technopolis 2010).

If the report is correct in that there is a "growing interest in the notion of research impact, evident amongst all funders" (Technopolis 2010), then the best way to avoid the irony experienced by participating institutions and researchers in the REF pilot exercise, is for researchers to not wait until a similar exercise is forced upon them.

Perhaps the ongoing-experiences of the HEFCE and the institutions that it covers, along with the experiences of programs such as 'Evaluating Research in Context' (ERiC) in the Netherlands (Spaapen *et al.*

2007), should be used as a basis to create a set of "best practices" for use by researchers not yet covered by any practical-impact requirements to do self-assessments, and potentially by funding institutions as a starting point for including practical impact in their evaluation of research.

## Conclusion

Since institutions have limited resources they must regularly assess if those resources are allocated efficiently, in regard to meeting the institution's goals. Traditionally, the focus of such evaluation has been on research quality using bibliometric methods that analyze article citations in academic journals.

More recently, there has been interest in the practical impact of research, that is, the benefit outside of research institutions. To evaluate the practical impact of research, a qualitative, case-study-based approach is typically recommended and/or used, ideally having the evaluation done by a panel of experts that includes the end-users of research.

The value of research comes from both its research quality and its practical impact. The best way to judge the value of research is to have the evaluation done by a panel of experts, using both quantitative and qualitative methods.

---

## References


- Adie, Euan, and William Roe. 2013. "Altmetric: Enriching Scholarly Content with Article-Level Discussion and Metrics." *Learned Publishing* 26 (1): 11–17. doi:10.1087/20130103.
- Adler, Robert, John Ewing, and Peter Taylor. 2009. "Citation Statistics." *Statistical Science* 24 (1): 1–14. doi:10.1214/09-STS285.
- Archambault, Éric, and Vincent Larivière. 2009. "History of the Journal Impact Factor: Contingencies and Consequences." *Scientometrics* 79 (3): 635–49. doi:10.1007/s11192-007-2036-x.
- Bollen, Johan, Marko A. Rodriquez, and Herbert Van de Sompel. 2006. "Journal Status." *Scientometrics* 69 (3): 669–87. doi:10.1007/s11192-006-0176-z.
- Bonacich, Phillip. 1987. "Power and Centrality: A Family of Measures." *American Journal of Sociology* 92 (5): 1170–82. doi:10.2307/2780000.
- Bornmann, Lutz, Werner Marx, Armen Yuri Gasparyan, and George D. Kitas. 2012. "Diversity, Value and Limitations of the Journal Impact Factor and Alternative Metrics." *Rheumatology International* 32 (7): 1861–67. doi:10.1007/s00296-011-2276-1.
- Broadus, R. N. 1987. "Toward a Definition of 'bibliometrics.'" *Scientometrics* 12 (5-6): 373–79. doi:10.1007/BF02016680.
- Butler, L. 2008. "Using a Balanced Approach to Bibliometrics: Quantitative Performance Measures in the Australian Research Quality Framework." *Ethics in Science and Environmental Politics* 8 (June): 83–92. doi:10.3354/esep00077.
- Connor, James. 2011. "Google Scholar Citations Open To All." *Google Scholar Blog*. Accessed April 12. http://googlescholar.blogspot.com/2011/11/google-scholar-citations-open-to-all.html.
- Donovan, Claire. 2008. "The Australian Research Quality Framework: A Live Experiment in Capturing the Social, Economic, Environmental, and Cultural Returns of Publicly Funded Research." *New Directions for Evaluation* 2008 (118): 47–60. doi:10.1002/ev.260.
- Egghe, L., and R. Rousseau. 1990. *Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science*. Amsterdam ; New York: Elsevier Science Publishers. http://catalog.hathitrust.org/Record/002225028.
- Egghe, Leo. 2013. "Theory and Practise of the G-Index." *Scientometrics* 69 (1): 131–52. doi:10.1007/s11192-006-0144-7.

- Elsevier. 2015. "Journal Metrics: Research Analytics Redefined." *Journal Metrics: Research Analytics Redefined*. Accessed April 7. http://www.journalmetrics.com/.
- Fleischmann, Martin, and Stanley Pons. 1989. "Electrochemically Induced Nuclear Fusion of Deuterium." *Journal of Electroanalytical Chemistry and Interfacial Electrochemistry* 261 (2): 301–8. doi:10.1016/0022-0728(89)80006-3.
- Garfield, Eugene. 1979. *Citation Indexing - Its Theory and Application in Science, Technology, and Humanities*. New York: Wiley.
- González-Pereira, Borja, Vicente P. Guerrero-Bote, and Félix Moya-Anegón. 2010. "A New Approach to the Metric of Journals' Scientific Prestige: The SJR Indicator." *Journal of Informetrics* 4 (3): 379–91. doi:10.1016/j.joi.2010.03.002.
- Grant, J., P. B. Brutscheer, S. Kirk, L. Butler, and S. Wooding. 2010. "Documented Briefing: Capturing Research Impacts–a Review of International Practice". *DB-578-HEFCE*. RAND Documented Briefings, RAND Corporation. http://www.rand.org/pubs/documented_briefings/DB578.html. Retrieved November 4, 2014. ._.
- Gregory, Jennie. 1937. "An Evaluation of Medical Periodicals." *Bulletin of the Medical Library Association* 25 (3): 172–88.
- Gross, P. L. K., and E. M. Gross. 1927. "College Libraries and Chemical Education." *Science* 66 (1713): 385–89.
- HEFCE. 2009a. *Research Excellence Framework - Second Consultation on the Assessment and Funding of Research.* HEFCE 2009/38. HEFCE. http://webarchive.nationalarchives.gov.uk/20100202100434/http://www.hefce.ac.uk/pubs/hefce/2009/09_38/09_38.pdf. Retrieved April 3, 2015.
- HEFCE. 2009b. *Report on the Pilot Exercise to Develop Bibliometric Indicators for the Research Excellence Framework.* HEFCE-REF 2009/39. Higher Education Funding Council for England (HEFCE). http://webarchive.nationalarchives.gov.uk/20100202100434/http://www.hefce.ac.uk/pubs/year/2009/200939/.
- HEFCE. 2015. "Our Role." Higher Education Funding Council for England. Accessed April 3. https://www.hefce.ac.uk/about/role/.
- Hackh, Ingo. 1936. "The Periodicals Useful in the Dental Library." *Bulletin of the Medical Library Association* 25 (1-2): 109–12.
- Hirsch, J. E. 2005. "An Index to Quantify an Individual's Scientific Research Output." *Proceedings of the National Academy of Sciences of the United States of America* 102 (46): 16569–72. doi:10.1073/pnas.0507655102.
- Hood, William W., and Concepción S. Wilson. 2001. "The Literature of Bibliometrics, Scientometrics, and Informetrics." *Scientometrics* 52 (2): 291–314. doi:10.1023/A:1017919924342.
- Ingwersen, Peter, and Finn Hjortgaard Christensen. 1997. "Data Set Isolation for Bibliometric Online Analyses of Research Publications: Fundamental Methodological Issues." *Journal of the American Society for Information Science* 48 (3): 205–17. doi:10.1002/(SICI)1097-4571(199703)48:3<205::AID-ASI3>3.0.CO;2-0
- Khazragui, Hanan, and John Hudson. 2015. "Measuring the Benefits of University Research: Impact and the REF in the UK." *Research Evaluation* 24 (1): 51–62. doi:10.1093/reseval/rvu028.
- Kostoff, R. N. 1998. "The Use and Misuse of Citation Analysis in Research Evaluation." *Scientometrics* 43 (1): 27–43. doi:10.1007/BF02458392.
- Lehmann, Sune, Andrew D. Jackson, and Benny E. Lautrup. 2006. "Measures for Measures." *Nature* 444 (7122): 1003–4. doi:10.1038/4441003a.
- MacRoberts, Michael H., and Barbara R. MacRoberts. 1989. "Problems of Citation Analysis: A Critical Review." *Journal of the American Society for Information Science* 40 (5): 342–49. doi:10.1002/(SICI)1097-4571(198909)40:5<342::AID-ASI7>3.0.CO;2-U.
- Martyn, John, and Alan Gilchrist. 1968. *An Evaluation of British Scientific Journals*. London: Aslib.
- Moed, Henk F. 2005. *Citation Analysis in Research Evaluation*. Dordrecht; [Great Britain]: Springer.

- Moed, Henk F. 2009. "New Developments in the Use of Citation Analysis in Research Evaluation." *Archivum Immunologiae et Therapiae Experimentalis* 57 (1): 13–18. doi:10.1007/s00005-009-0001-5.
- Moed, Henk F. 2010. "Measuring Contextual Citation Impact of Scientific Journals." *Journal of Informetrics* 4 (3): 265–77. doi:10.1016/j.joi.2010.01.002.
- Moed, Henk F. 2011. "The Source Normalized Impact per Paper Is a Valid and Sophisticated Indicator of Journal Citation Impact." *Journal of the American Society for Information Science and Technology* 62 (1): 211–13. doi:10.1002/asi.21424.
- Mohammadi, Ehsan, Mike Thelwall, Stefanie Haustein, and Vincent Larivière. 2014. "Who Reads Research Articles? An Altmetrics Analysis of Mendeley User Categories." *Journal of the Association for Information Science and Technology*, 1–27.
- Schreiber, Michael. 2008. "A Modification of the H-Index: The Hm-Index Accounts for Multi-Authored Manuscripts." *Journal of Informetrics* 2 (3): 211–16. doi:10.1016/j.joi.2008.05.001.
- Spaapen, J.B, H Dijstelbloem, and F.J.M Wamelink. 2007. *Evaluating Research in Context: A Method for Comprehensive Assessment.* The Hague: Consultative Committee of Sector Councils for Research and Development (COS).
- Tague-Sutcliffe, Jean. 1992. "An Introduction to Informetrics." *Information Processing & Management* 28 (1): 1–3. doi:10.1016/0306-4573(92)90087-G.
- Technopolis Ltd. 2010. *REF Research Impact Pilot Exercise Lessons-Learned Project: Feedback on Pilot Submissions.* Higher Education Funding Council for England. http://www.ref.ac.uk/pubs/refimpactpilotlessons-learnedfeedbackonpilotsubmissions/.
- West, Jevin D. 2015. "Eigenfactor." *Eigenfactor.* Accessed April 9. http://www.eigenfactor.org/methods.php.
- Zahedi, Zohreh, Rodrigo Costas, and Paul Wouters. 2014. "How Well Developed Are Altmetrics? A Cross-Disciplinary Analysis of the Presence of 'alternative Metrics' in Scientific Publications." *Scientometrics* 101 (2): 1491–1513. doi:10.1007/s11192-014-1264-0.
- van Eck, Nees Jan, Ludo Waltman, Anthony F. J. van Raan, Robert J. M. Klautz, and Wilco C. Peul. 2013. "Citation Analysis May Severely Underestimate the Impact of Clinical Research as Compared to Basic Research." *PloS One* 8 (4): e62395. doi:10.1371/journal.pone.0062395.
- van Leeuwen, Thed N., H. F. Moed, and J. Reedijk. 1999. "Critical Comments on Institute for Scientific Information Impact Factors: A Sample of Inorganic Molecular Chemistry Journals." *Journal of Information Science* 25 (6): 489–98. doi:10.1177/016555159902500605.
- van Raan, A., H. Moed, and T. van Leeuwen. 2007. *Scoping Study on the Use of Bibliometric Analysis to Measure the Quality of Research in UK Higher Education Institutions.* HEFCE 2007/34. http://webarchive.nationalarchives.gov.uk/20120118171947/http://www.hefce.ac.uk/pubs/rdreports/2007/rd18_07/. Retrieved February 8, 2015.