Alternative metric indicators for funding scheme evaluations

Mike Thelwall and Kayvan Kousha
University of Wolverhampton, Wolverhampton, UK, and
Adam Dinsmore and Kevin Dolby
Wellcome Trust, London, UK

Received 15 September 2015
Revised 21 October 2015
Accepted 26 October 2015

Abstract

Purpose – The purpose of this paper is to investigate the potential of altmetric and webometric indicators to aid with funding agencies' evaluations of their funding schemes.
Design/methodology/approach – This paper analyses a range of altmetric and webometric indicators in terms of suitability for funding scheme evaluations, compares them to traditional indicators and reports some statistics derived from a pilot study with Wellcome Trust-associated publications.
Findings – Some alternative indicators have advantages to usefully complement scientometric data by reflecting a different type of impact or through being available before citation data.
Research limitations/implications – The empirical part of the results is based on a single case study and does not give statistical evidence for the added value of any of the indicators.
Practical implications – A few selected alternative indicators can be used by funding agencies as part of their funding scheme evaluations if they are processed in ways that enable comparisons between data sets. Their evidence value is only weak, however.
Originality/value – This is the first analysis of altmetrics or webometrics from a funding scheme evaluation perspective.

Keywords Altmetrics, Research evaluation, Funding programme, Funding scheme, Funding stream, Webometrics
Paper type Research paper

Introduction

Large funding agencies evaluate the effectiveness of their funding decisions, and the impact of the work they fund, through a combination of qualitative methodologies (e.g. surveys, focus groups) and quantitative evaluations (e.g. citation analysis). For example, an organisation may wish to compare the performance of their different funding schemes in order to judge whether its present strategy is producing the maximum impact on its investment. They may also evaluate new funding schemes to inform decisions regarding whether they should stop, continue or expand.

Qualitative assessments are probably the only realistic choice for assessing individual projects; quantitative indicators are not reliable for small amounts of research as they cannot capture the overall context of how the research was conducted and the true impact it may have had. If individual project assessments are reviewed and graded in a way that would allow aggregate results to be compared or summarised (e.g. Hamilton, 2011),

Thank you to Jonathan Levitt and Chonnettia Jones for comments on an earlier version of this paper. Though no specific funding was awarded in support of this paper, KD and AD are employed by the Wellcome Trust.

Aslib Journal of Information Management
Vol. 68 No. 1, 2016
pp. 2-18
Emerald Group Publishing Limited
2050-3806
DOI 10.1108/AJIM-09-2015-0146

then it could be straightforward to reuse the results for funding stream evaluations. Nevertheless, the subjective nature of qualitative assessments and the potential for different criteria being used in different contexts may undermine such attempts at aggregation. For example, one interim programme evaluation taking this approach reported that, "62% of services state that independent review or other similar evidence indicate that a majority or nearly all research is world-leading in terms of its originality, significance and rigor" (Annerberg *et al.*, 2010, p. 60). Even though this is apparently high-quality evidence, it is probably difficult to ensure that assessments of individual funded projects are sufficiently critical to be credible. In particular, there may be disincentives to giving low scores, such as loss of goodwill from the peers evaluated if the evaluation is not anonymous, and perhaps even a loss of face to the funding organisation or sections of it, if they appear to have funded poor research. Moreover, individual project evaluations may be reported in ways that are not conducive to effective scheme evaluations, such as if a simple pass/fail grade is given. Evaluations may also rely upon interviews or self-reported information from the awardees (e.g. Meagher, 2009), which may not fully reveal any weaknesses. Alternatively, panels of experts may be convened to assess a particular funding scheme perhaps after presentations and interviews with grant holders (e.g. EPSRC, 2011).

Evaluations can also draw upon indicators of research quality from the process of awarding funding. For example, the interim evaluation of the European Union seventh framework programme drew upon evidence of the grades given to the submitted proposals as evidence of the health of the initiative from the perspective of the quality of the formulation of the submitted research projects (Annerberg *et al.*, 2010). This is also evidence of the quality of the funded projects since because only those with the highest peer review scores are funded. A wider issue that is not considered here is the extent to which the funding has improved the performance of those funded, assuming that they would have continued to research anyway. This is probably a more important issue for non-academic research (Jaffe, 2002).

Another common strategy for evaluating research schemes is to assess the number and citation impact of the project outputs. This has the advantage of being transparent, independent and giving results that can be compared to bibliometric benchmarks or between schemes and funding agencies, if appropriately normalised. Research projects typically have to report a list of their publications and so funding agencies tend to have the raw data to analyse publications, although they may need to collect appropriate citation data. Bibliometric indicators have known biases, however. For example, citation-based impact indicators typically reflect an aspect of academic impact whereas funding agencies often target wider types of impact. Moreover, whereas citation impact can be benchmarked against world norms, this may not be possible for productivity metrics and the overall volume of research produced seems to be important as well as its excellence. A second problem with citation-based indicators is that they impose a delay of several years on evaluations. Citations take time to accrue whilst publications are read by others who then conduct new research informed by their reading, write up and publish studies and then wait for them to be peer reviewed and published. Alternative indicators derived from the web may be timelier because much of the web and all of the social web allows instant publishing. Web-derived indicators (for a review, see Thelwall and Kousha, 2015a, b; Kousha and Thelwall, 2015) may also reflect wider impacts than citation counts because non-academics widely publish online and may sometimes discuss

researchers and research-related issues (Cronin *et al.*, 1998; Priem *et al.*, 2010). The web may also contain evidence of educational uses of research, such as citations in online syllabi (Kousha and Thelwall, 2008). This study explores a variety of alternative indicators, provides guidelines for research funding scheme evaluations and reports a pilot study of the Webometric/altmetric impact of research outputs produced by a sample of researchers supported by the Wellcome Trust.

Altmetric and webometric indicators for funding scheme evaluations

A funding scheme typically publishes a set of eligibility criteria and the amount and type of funding available, and then initiates a competitive submission process where peer reviews of the proposals contribute to the selection of a subset to award grants. Funding schemes can have widely different general goals, such as to drive forward science or to generate societal impacts from research, as well as a broad or narrow topic focus. They may also have secondary goals that can affect eligibility criteria, such as encouraging early career researchers or international collaboration. All of these affect the types of outputs that can be expected from the funded research as well as the criteria with which they should be evaluated. For example, the impacts of a research scheme with an unrestricted applicant pool would presumably tend to be greater than the impacts of an early career researcher scheme. When bibliometric or other quantitative data is gathered, then, it should be benchmarked against comparable data whenever possible (such as the same funding scheme for the previous year, or for another funding agency) or the significance of any differences found should be evaluated qualitatively, which may be difficult. When using any type of impact indicator the following are needed:

- A clear indication of the type of impact reflected by the indicator. This should be based upon evidence of use in context, rather than user base. For example, although a tiny minority of Twitter users are academics, the majority of tweets about academic articles are probably written by academics. This is most important when only a few indicators are used, when evidence of a specific type of impact is needed, or when a range of highly related indicators are used.
- A rationale for using the indicator. For example, this could be to reflect a different type of impact to that of other available indicators, to give earlier evidence of impact, or for triangulation with other indicators. If used primarily for triangulation, then the indicator should not be substantially less robust than the indicators that it is triangulated with.
- A narrow enough citation window so that the window itself does not have a big impact on the scores, by older articles being more highly cited because they are older rather than because they tend to have had more impact. A year is reasonable for citation counts and most altmetrics, except perhaps for articles published in the previous two years, and half a year is more reasonable for social web indicators with a fast increasing uptake, such as Twitter. For some metrics, explicit citation windows can be set but for others, the indicator cannot be calculated over a user-specified period of time (e.g. citations only from articles published 2012-2013). In the latter case, the cited articles should only be compared against other articles published at approximately the same time (e.g. year or half-year).

- Sufficient coverage of publications so that statistics derived from them can reflect the impact of typical outputs, if this is desired. For example, an indicator with zero values for 99 per cent of a funding scheme's outputs is unlikely to be informative about its typical impacts. However, for very large collections of publications, statistical averaging may allow indicators with low coverage to be useful as indicators of the impacts for a minority of outputs.
- A high-enough ratio of value to randomness and bias for the indicators to be reliable on the scale at which they are used (relates to the above). Studies showing at least a moderate correlation between the indicator and another metric of better known value would be evidence of a sufficient ratio of value to randomness. A standard way to indicate reliability is to calculate confidence intervals for the mean. Since citation and altmetric data are often highly skewed and with many zeros, confidence intervals may be easiest to calculate indirectly from the logarithm of one plus the data (e.g. Thelwall and Wilson, 2014) in order to use confidence intervals designed for the normal distribution. Alternatively, statistical hypothesis tests could be conducted for the significance of any important differences.

Alternative indicators for research assessment

Citation counts are widely used in many types of research assessments, supported by evidence that appropriately normalised indicators correlate with peer review judgements in many fields (Franceschet and Costantini, 2011; Nederhof and Van Raan, 1993; Rinia *et al.*, 1998) and other indicators are also used to cover their limitations. Appropriately field and year normalised citation counts seem to be reasonable indicators of the scholarly impact of academic articles. Field normalisation is needed because different research areas naturally attract different average numbers of citations and time normalisation is needed because older articles tend to be more cited. Aggregation over a sufficiently large set is also important because articles can make big contributions to science without being highly cited, for example, by closing off invalid areas of research or by finding a definitive solution to a long-term problem. Individual articles can also be highly cited to be criticised or for a useful review of a field without contributing significant novelty. For example, the Fleischmann-Pons cold fusion article has 441 Scopus citations, including some from papers failing to replicate its results and others arguing against its results. It is now widely thought to be erroneous science (Taubes, 1993). Over a reasonably large set of documents, these variations should tend to cancel out, especially if the document set evaluated is relatively homogeneous (e.g. covering only refereed journal articles and excluding review articles because these tend to be more highly cited as a genre, but contain no primary research). Nevertheless, the field normalisation is likely to use simplistic heuristics because science is inherently difficult to classify and is increasingly multidisciplinary. This may cause systematic biases if two or more groups are compared that tend to specialise in subfields with differing citation norms. As a result of these limitations, citation metrics are often used to inform rather than replace peer review for important evaluations, such as for several sub-panels in the UK REF2014 (2012). They are partly used to replace peer review in some important evaluations, however, such as for specified subsets of articles in the Italian research evaluation exercise (Abramo and D'Angelo, 2015).

Since citations can only reflect academic impact, other metrics have been developed to reflect different types of impact. Patent metrics (Meyer, 2003; Narin, 1994;

Oppenheim, 2000; Tijssen, 2001) are an obvious choice: the number of patents granted could be seen as an indicator of the quantity of commercially relevant research produced, income from patents might reflect the commercial value of the research generating them, or citations from patents to traditional academic publications might reflect the commercial relevance of the cited work. Patent citations are used in some major research evaluations, such as Excellence in Research Australia (ERA, 2010), but are limited in scope because the extent to which inventions are patented varies extensively by commercial sector, with many businesses preferring secrecy instead (Cohen *et al.*, 2000).

Alternative metrics derived from the web (Almind and Ingwersen, 1997), online scholarly sources (Neylon and Wu, 2009) and particularly the social web (Priem *et al.*, 2010), have also been proposed as indicators for academic research. The idea here is that the web is not just used by academics and therefore data from the web about academic research may be useful as evidence of the wider impacts of that research. No indicator can be taken at face value, however, and each needs to be assessed for its potential. Most indicator assessments have correlated them against citation counts in order to demonstrate that they relate to academic impact in some way and that they are not completely random (Sud and Thelwall, 2014). A few have also focused on identifying the type of wider impact that the indicators reflect, if any, and some have investigated their growth over time. Here is a brief summary of some key alternative webometrics and altmetrics for academic articles. See the cited references for information about how to calculate each one:

- PowerPoint file mentions: these are apparently an indicator of educational and scholarly impact through the posting of teaching files online as well as research presentations and have a low, statistically significant positive correlation with citation counts but are rare (Thelwall and Kousha, 2008).
- PDF and DOC mentions: mentions from PDF and Microsoft Word documents online may be a useful indicator of online grey literature citation impact (Wilkinson *et al.*, 2014), although there is no evidence yet about the prevalence of PDF and DOC mentions across scientific fields, nor about the magnitude of their correlation with citation counts.
- Web (course) syllabus mentions: these are an indicator of educational impact (Kousha and Thelwall, 2008) and are more common as citations in the social sciences, where their correlation with citations may be about 0.2, but are almost non-existent in the natural and life sciences. Citations from online academic syllabi are also a potential indicator of educational value for books and monographs in the social sciences and humanities (Kousha and Thelwall, in press).
- Web mentions: counting mentions of an academic publication on the web turns the web into a citation database and allows citations to be captured from non-academic sources. It is not useful in practice because articles with short titles are hard to search for online and all articles may appear in numerous routine library lists and academic CVs (Vaughan and Shaw, 2003).
- URL citations: counting mentions of the URLs of academic publications on the web is also not useful in practice because articles have multiple URLs, often in a complex format, and the URLs may appear in library lists (Kousha and Thelwall, 2007).

- Tweet links: Twitter is very widely used in many countries (China is an important exception) both by academics and the public and there are probably more tweets about recent academic articles than counts for any other altmetric (Thelwall *et al.*, 2013). It seems that tweets are an indicator of interest in an article by other academics (Thelwall *et al.*, 2013) and their association with citation counts is mostly very weak (Eysenbach, 2011; Shuai *et al.*, 2012; Haustein *et al.*, 2014; Thelwall *et al.*, 2013) so there seems to be little value in using them even for academic impact evidence.
- Blog citations: although blog citations of academic articles correlate with academic impact (Shema *et al.*, 2014), they seem to be much too rare and difficult to identify to be worth systematically gathering for research assessments.
- F1000 scores: these post-publication peer-review evaluations can be informative indicators of scientific and non-scientific value for biomedical science articles (Li and Thelwall, 2012; Mohammadi and Thelwall, 2013), although only a minority have scores and they may need to be bought from F1000. Since 90 per cent of these appear within half a year of an article being published (Waltman and Costas, 2014), they have a substantial time advantage over citations.
- Online clinical guideline citations: citations from online clinical guidelines are direct evidence of the health benefits of medical research and these citations correlate weakly with citation counts (Thelwall and Maflahi, in press). Probably a very small proportion of medical articles are cited by guidelines, however, and they are time consuming to identify online and so are probably not yet a practical data source. The addition of some guidelines (NICE, WHO) to the NLM Bookshelf (www.ncbi.nlm.nih.gov/books), with links to PubMed records is starting to make this easier.
- Mendeley readers: the number of registered users of the social reference sharing site Mendeley that bookmark an article correlates highly (about 0.7) with citation counts in many fields and contexts (Li *et al.*, 2012; Thelwall and Wilson, in press), and seem to be more prevalent than other altmetrics (Zahedi *et al.*, 2014a) except perhaps tweet counts (see also Borrego and Fry, 2012). These readership counts mainly reflect academic impact, although with an element of educational impact and a bias towards younger researchers (Mohammadi *et al.*, 2015, in press; see also Zahedi *et al.*, 2014b). Mendeley readers typically appear about one to two years before citations (Maflahi and Thelwall, in press), making them particularly useful for early impact evaluations.

In addition to the above, Facebook wall posts, Zotero and CiteULike bookmarks, Reddits and LinkedIn citations seem to be too rare for use in evaluations, except perhaps those on a very large scale. There is some evidence of a weak correlation with citation counts for most of these (Costas *et al.*, 2015; Thelwall *et al.*, 2013).

A generic problem with altmetrics and webometrics is that they are typically easy to manipulate because they are not subject to quality control. They are not suitable for formal evaluations of researchers or research groups (Wouters and Costas, 2012) unless steps are taken to guard against deliberate fraud. This is probably not a concern for funding scheme analyses, however, assuming that these would not have stakeholders with sufficient interest to risk systematically manipulating any alternative indicators used (see also Colquhoun and Plested, 2014). For example, a funding scheme analysis may cover hundreds of researchers, and it seems unlikely that any would feel strongly enough about the continuation of a new funding scheme and knowledgeable enough

about evaluation policies to risk manipulating the data used for them on a large enough scale to have an influence. Social web indicators are probably biased against impacts of more senior scholars, who may be less likely to use the new sites (Mas-Bleda *et al.*, 2014).

Wellcome Trust pilot study

The Wellcome Trust in the UK is one of the world's largest private funders of medical research and operates a number of funding schemes for different career stages and subject areas. The Wellcome Trust Insight and Analysis Team have been exploring the potential value of altmetrics and webometrics to research funding assessment for some time (Dinsmore *et al.*, 2014) and in January 2015 compiled a data set combining citation, altmetric and Webometric data to facilitate a pilot study of their properties.

Accurately matching publication outputs to specific grants is not straightforward: despite clear guidance from the Wellcome Trust on how to acknowledge funding which includes the unique grant number (www.wellcome.ac.uk/Managing-a-grant/End-of-a-grant/WTD037950.htm), many papers are published with only vague mentions of the funding source. Different manuscript submission process and publisher requirements are likely to be linked to these problems, and with some papers having dozens of contributing authors it is likely that the corresponding author just does not have accurate funding details for all their colleagues. The data set analysed here consists of 5,087 academic publications from 2007 to 2013 that have been linked to specific Wellcome Trust grants, and hence funding schemes, through the use of unique grant numbers recorded in the Web of Science, or through periodic searches of Web of Science data to identify papers authored by researchers receiving personal awards from the Trust and manual checking of matching records. This is an ongoing process as part of Wellcome's monitoring of the outcomes of its awards. The data set contained some duplicates due to outputs being associated with more than one grant. Because of likely inaccuracies and under-counting in the data due to the funding acknowledgement issues, the individual schemes have been anonymised for this pilot exercise, but can be broken down as follows:

- Scheme 1: 394;
- Scheme 2: 775;
- Scheme 3: 562;
- Scheme 4: 843;
- Scheme 5: 581; and
- Scheme 6: 1,833.

Another reason for anonymising the schemes is to ensure non-prejudicial judgements of the data given that schemes may be expected to have differing levels of impacts amongst different audiences.

Most projects are also classified into funding streams which represent the broad subject area of the initial application (although this is largely for internal Wellcome classification and may not accurately line up with the subject matter of resulting publications). The streams with at least 100 articles are listed below, together with the most common associated scheme:

- infection and immuno-biology (1,072 publications; including 508 from Scheme 6);
- neuroscience and mental health (994; 306 Scheme 6);
- population health (748; 317 Scheme 6);

- genetic and molecular sciences (622; 356 Scheme 6);
- cellular, developmental and physiological sciences (500; 319 Scheme 6);
- molecules, genes and cells (194; 167 Scheme 4); and
- medical humanities (116).

One of the issues that the Wellcome Insight and Analysis Team would like to address is whether certain funding streams or schemes are generating attention and potential impact amongst different audiences. In theory, indicators of impact could help such an evaluation by enabling numerical comparisons between streams. A variety of types of data were collected for this. Altmetric data were mostly supplied by altmetric.com (Adie and Roe, 2013), including the following:

- F1000 reviews: number of the article in F1000, as reported by altmetric.com in January 2015;
- Tweeters: number of Twitter accounts containing a link to the article, as reported by altmetric.com in January 2015;
- News outlets: number of news stories containing a link to the article, as reported by altmetric.com in January 2015;
- CiteULike readers: number of CiteULike readers of the article, as reported by altmetric.com in January 2015;
- Facebook walls: number of Facebook wall posts containing a link to the article, as reported by altmetric.com in January 2015;
- Bloggers: number of Bloggers with a post containing a link to the article, as reported by altmetric.com in January 2015;
- Reddit threads: number of Reddit threads containing a link to the article, as reported by altmetric.com in January 2015;
- Altmetric score: overall (composite) Altmetric.com score for the article, as reported by altmetric.com in January 2015; and
- Mendeley readers: number of Mendeley readers of the article, as collected by Webometric Analyst in June 2014.

The following webometric data were collected from the Web using Webometric Analyst:

- Syllabus (PDF/DOC): number of PDF or Word documents online matching a Bing search for online syllabi mentioning the article, as collected in June 2014;
- Google Books cites: number of Google Book search citations to the article, as collected in June 2014 (see also Kousha and Thelwall, 2014);
- PDF mentions: number of PDF documents matching a Bing search for the article, as collected in June 2014; and
- DOC mentions: number of Word documents matching a Bing search for the article, as collected on in June 2014.

The following citation data derived from the Web of Science was also used:

- Total cites (end 2013): total number of citations to the article, as recorded in the Web of Science at the end of 2013; and

- JIF 2013: the Thomson Reuters Journal Impact Factor (JIF) 2013 for the journal publishing the article.

The indicators collected above may relate to different types of impact. The Web of Science data and Google Books citations presumably reflect an aspect of scholarly impact (Kousha and Thelwall, 2014; Moed, 2006). The CiteULike readers and Mendeley readers probably reflect mainly academic impact but also some educational impact, particularly from graduate students (Mohammadi *et al.*, 2015). In contrast, syllabus mentions specifically indicate educational impact (Kousha and Thelwall, 2008). In addition, news outlets probably reflect public interest in a topic, and F1000 reviews probably reflect mainly academic interest but also partly value for practitioners (Mohammadi and Thelwall, 2013). Both PDF and DOC mentions probably reflect a combination of academic, education and organisational (i.e. via white papers) impact. The remaining data sources are open to all users, however, and so the type of impact that they reflect, if any, depends upon which groups of users, if any, predominantly discuss research, and why. For example, if most people blogging about science are PhD students or academics then blog citations could reflect academic impact, unless they blogged to attract a public outreach goal (Shema *et al.*, 2015), in which case blog citations may be conceptually similar to news citations. It seems that tweet citations are the only social web indicator for which there is evidence of the type of impact they tend to indicate, which is predominantly academic, but with a bias towards publicity (Thelwall *et al.*, 2013). In other words, although the vast majority of Twitter users are non-scholars, the minority that tweet about research are mainly online scholars. For the other indicators, little is known about who uses them to cite academic research and why.

For a fair analysis, only articles from the same year should be analysed and so a focus was given to the 2,625 articles from 2012. The choice of 2012 was due to comprehensive altmetric.com data being available only for 2012 and 2013 publications and the earlier of these two years had the largest counts for most of the metrics.

A variety of practical considerations related to the time at which the publication lists were built had resulted in incomplete data for most of the articles. For a fair comparison, articles for which at least one of the metrics was unavailable were therefore removed, leaving 1,467. The altmetric data for all articles from 2012 correlated positively and significantly with Scopus citation counts, varying from 0.117 (F1000 reviews) to 0.777 (Mendeley readers). The webometric data for all articles from 2012 also correlated positively and significantly with Scopus citation counts, varying from 0.328 (DOC mentions) to 0.591 (Google Books cites). The lower correlations could be due to there being little data for the indicator, the data originating from multiple fields (Thelwall and Fairclough, 2015) or the data source being noisy, all of which are technical limitations. Alternatively, lower correlations could also be caused by an indicator reflecting a non-academic type of impact, which is an advantage in a research evaluation context. Probably all of these reasons apply to some extent for the indicators.

Funding organisations sometimes compare different funding schemes or streams as well as the same scheme or stream in different years. The purpose of this is to identify particularly successful strategies that may be more fully funded in future, as well as to assess whether new schemes are generating the expected level of impact. Hence, an important role for citations and alternative indicators is to help to assess the relative merits of different funding schemes. This paper reports one such comparison as a proof of concept.

In order to avoid distortions in a comparison between schemes, articles from different fields should not be compared without normalisation. Since normalisation data were not available for the webometric and altmetric data, streams (fields) needed

to be analysed separately. The largest stream in 2012 was Neuroscience and Mental Health, and the three schemes with the most articles in this scheme were Scheme 6 (61), Scheme 3 (46) and Scheme 5 (45). These were chosen for a comparison as a purely pragmatic step due to the amount of data, rather than any specific Wellcome Trust policy need. To compare the metrics Scheme 6 was compared with the other two schemes with the assumption that a more effective metric would detect a difference. Hence, lower p-values for an independent samples t-test for the difference between the two suggest more powerful metrics. Since most citation-like data is skewed, all indicators (except the JIF and F1000 scores) were transformed with the formula $\ln(1+x)$ in order to make it reasonable to apply statistics based on the normal distribution (e.g. see Thelwall and Wilson, 2014). The test is biased against citations because these were collected at an earlier period of time. The results are therefore only a proof of concept rather than evidence of the relative values of the different metrics (Tables I and II).

The geometric mean is a more precise than the arithmetic mean for highly skewed data, as is typical for citations and many altmetrics, and so this was used as the default measure of average in the tables (Fairclough and Thelwall, 2015a; Zitt, 2012).

Table I suggests that Scheme 6 has an impact advantage over Scheme 3 because the means tend to be higher for the indicators with low p-values. No systematic pattern is evident in Table II, however. Hence, although we can make no claims to the true impact of the research produced by the two schemes, it is reasonable to say that Scheme 6 produces research that receives more online attention than research produced by Scheme 3, at least for the Neuroscience and Mental Health category and on some of the metrics. Nevertheless, the p-values cannot be relied upon because false positives are to be expected when a large set of tests are conducted simultaneously, necessitating a Bonferroni correction or a more careful analysis of the results. Another problem is that a paper may be the result of funding from multiple schemes and so the data sets of papers linked to each scheme are not properly independent.

A Bonferroni correction for $n = 14$ would correct the p-value from 0.05 to 0.004 and with this correction the only statistically significant result is for F1000 reviews in Table II.

Indicator	p	Scheme 5 mean	Scheme 6 mean
Tweeters	0.027	0.73	1.10
Altmetric score	0.031	0.76	1.13
Syllabus (PDF/DOC)	0.045	0.00	0.02
Facebook walls	0.048	0.07	0.16
Reddit threads	0.050	0.00	0.02
Google Books cites	0.091	0.28	0.17
Bloggers	0.116	0.04	0.10
Total cites (end 2013)	0.158	1.89	1.57
JIF 2013[a]	0.317	8.34	6.97
Mendeley readers	0.333	3.19	3.58
PDF mentions	0.436	0.25	0.21
DOC mentions	0.465	0.08	0.06
News outlets	0.633	0.03	0.05
CiteULike readers	0.746	0.22	0.21
F1000 reviews[a]	0.974	0.04	0.04

Notes: The data are Wellcome-funded refereed journal articles published in 2012 and categorised as neuroscience and mental health and within either Scheme 5 (45) or Scheme 6 (61). [a]Uses the standard mean rather than the geometric mean

Table I. Independent samples t-tests to compare Scheme 5 with Scheme 6, based on the geometric mean, for a variety of metrics

Indicator	p	Scheme 3 mean	Scheme 6 mean
F1000 reviews[a]	0.004	0.00	0.04
CiteULike readers	0.022	0.38	0.21
News outlets	0.037	0.01	0.05
Mendeley readers	0.062	4.30	3.58
Google Books cites	0.186	0.24	0.17
Facebook walls	0.191	0.10	0.16
PDF mentions	0.307	0.27	0.21
Pinterest posts	0.323	0.01	0.00
JIF 2013[a]	0.568	7.60	6.97
Reddit threads	0.576	0.03	0.02
Total cites (end 2013)	0.822	1.61	1.57
Tweeters	0.833	1.07	1.10
Altmetric score	0.896	1.11	1.13
Bloggers	0.915	0.10	0.10
Syllabus (PDF/DOC)	0.994	0.02	0.02

Notes: The data are Wellcome-funded refereed journal articles published in 2012 and categorised as neuroscience and mental health and within either Scheme 3 (46) or Scheme 6 (61). [a]Uses the standard mean rather than the geometric mean

Table II.
Independent samples *t*-tests to compare Scheme 3 with Scheme 6, based on the geometric mean, for a variety of metrics

This raw data for this result is essentially the fact that none of the Scheme 3 values had a F1000 review and 8 of the Scheme 6 values had a F1000 review. Hence, it seems that articles produced by Scheme 6 are significantly more likely to attract F1000 reviews than Scheme 3 articles. Out of the 8 F1000 articles, only two had authors in common and all articles were in different journals and so the difference is not due to an individual author or topic attracting attention from F1000. A possible explanation is that articles from one set of researchers are more likely to be picked up for reviewing by F1000, although the possible incompleteness of the data should again be noted.

Discussion and limitations

The results are subject to major limitations. Although some of the tests in Tables I and II above give statistically significant results, they do not prove that one funding scheme has had more impact of any particular type than the other. The significant differences may have been caused by many reasons other than differences in average impact between schemes. Perhaps most importantly, differences in the fields covered and methodological orientation of the schemes (e.g. pure or applied) can greatly affect the indicators generated. Moreover, the statistical tests assume that the data are independent whereas many are from the same authors or co-author teams since an individual grant may cover multiple years of research. It is unlikely that the impacts of the papers produced by an individual or group are independent of each other.

For the majority of indicators, another limitation is that they are not from quality controlled sources, such as academic publications, and therefore include an unknown amount of promotional material and spam. For example, one funding scheme may attract many tweets for its research because one of its large projects ran a particularly successful social media publicity initiative (e.g. Mollett *et al.*, 2011). Although many alternative indicators have previously been shown to correlate positively and significantly with citation counts, which gives evidence of their value, these correlations have mostly been weak (Thelwall and Kousha, 2015a, b; Kousha and Thelwall, 2015), often due to issues like the counts being too low to be useful (Thelwall *et al.*, 2013; Haustein *et al.*, 2014)

or insufficiently close connections with any kind of impact (Thelwall *et al.*, 2013). The main exceptions from the literature are counts of Mendeley readers (Fairclough and Thelwall, 2015b) and Google Books citations (Abdullah and Thelwall, 2014; Kousha and Thelwall, 2015), both of which seem to be numerous and high-quality sources. Of these, the former is the most promising for evaluations because of its ability to generate early impact evidence, whereas the latter is a slower moving source of evidence, although it might be useful for long-term evaluations in the humanities, where the cultural benefits of books may be important (Zuccala and Guns, 2013; but see also Hammarfelt, 2014; Thelwall and Delgado, 2015; Torres-Salinas *et al.*, 2012).

The above limitations are generic to the use of alternative indicators for research evaluation rather than specific to the case study. Whilst some of them could be ameliorated by systematic checking of data sources, such as reading tweets to identify spam, this may be too time consuming to be practical. Despite these limitations, however, the indicators may have value as a starting point for an evaluation. Whilst very weak indicators would probably be a distraction, the stronger indicators and the ones with a clearer interpretation could be helpful by suggesting a starting point for which funding streams have had particular types of strong impact. The evaluators should then assess the evidence and combine it with their own evaluations of the schemes in order to make a final judgement.

Conclusions and recommendations

This paper has discussed the potential value of altmetrics for funding scheme evaluations and given a small pilot test from the Wellcome Trust. There are a few alternative metrics that could give added value to funding scheme evaluations by providing indicators that suggest the average relative impacts of funding schemes to inform expert judgements. The most promising are Google Books citations, Mendeley readers and (for biomedical science), F1000 reviews. Nevertheless, for a fair comparison only articles from the same year and field should be compared and differences between the purposes of funding streams should be taken into account. The following are key recommendations for future funding stream evaluations and should be added to the bullet point list of recommendations above:

- Select a limited number of metrics to use for the evaluation rather than an exhaustive list. This is to avoid the increased uncertainty caused by carrying out multiple tests (the Bonferroni correction issue) in a way that weakens the power of each individual test. Nevertheless, the indicators selected should clearly be labelled as informative but not definitive and should be accompanied by caveats about their interpretation and robustness.
- The most promising alternative metrics seem to be F1000 scores for biomedical science, Google Books citations for humanities and book-oriented research and Mendeley readers for recent articles.
- Compare articles from the same field (and year) because citation and altmetric patterns vary between fields and years.
- Use a $\ln(1+x)$ transformation to allow more precise confidence intervals or simple statistical tests to be calculated to distinguish between the impacts of different funding schemes and report the geometric mean for all highly skewed indicators.
- Report confidence intervals for the (geometric) mean or conduct hypothesis tests to compare different funding schemes, including Bonferroni corrections for multiple simultaneous tests.

- Check positive results for the influence of individual authors or teams.
- Interpret the results in the context of the goals of the funding schemes – for example, some schemes will be expected to have mostly academic impact, while others are targeted more at translational or policy impacts.
- Take into account productivity and types of impact that are not reflected in the indicators used.

As the above recommendations suggest, considerable care needs to be taken with the collection, processing and interpretation of alternative indicators and the results are unlikely to give clear-cut conclusions. Nevertheless, these indicators may be useful for early impact evidence or to reflect alternative types of impact, particularly when large collections of publications are available to be assessed, but should be used to inform human judgements rather than to replace them. When given to the evaluators in this context, the data should be accompanied by caveats about its robustness and limitations so that it can perhaps be taken as a starting point for discussion but should not drive conclusions.

References

Abdullah, A. and Thelwall, M. (2014), "Can the impact of non-Western academic books be measured? An investigation of Google Books and Google Scholar for Malaysia", *Journal of the Association for Information Science and Technology*, Vol. 65 No. 12, pp. 2498-2508.

Abramo, G. and D'Angelo, C.A. (2015), "The VQR, Italy's second national research assessment: methodological failures and ranking distortions", *Journal of the Association for Information Science and Technology*, Vol. 66 No. 11, pp. 2202-2214.

Adie, E. and Roe, W. (2013), "Altmetric: enriching scholarly content with article-level discussion and metrics", *Learned Publishing*, Vol. 26 No. 1, pp. 11-17.

Almind, T.C. and Ingwersen, P. (1997), "Informetric analyses on the world wide web: methodological approaches to 'webometrics' ", *Journal of Documentation*, Vol. 53 No. 4, pp. 404-426.

Annerberg, R., Begg, I., Acheson, H., Borrás, S., Hallén, A., Maimets, T., Mustonen, R., Raffler, H., Swings, J.-P. and Ylihonko, C. (2010), *Interim Evaluation of the Seventh Framework Programme: Report of the Expert Group*, European Commission, available at: http://ec.europa.eu/research/evaluations/pdf/archive/other_reports_studies_and_documents/fp7_interim_evaluation_expert_group_report.pdf (accessed 30 November 2015).

Borrego, Á. and Fry, J. (2012), "Measuring researchers' use of scholarly information through social bookmarking data: a case study of BibSonomy", *Journal of Information Science*, Vol. 38 No. 3, pp. 297-308.

Cohen, W.M., Nelson, R.R. and Walsh, J.P. (2000), "Protecting their intellectual assets: appropriability conditions and why US manufacturing firms patent (or not)", NBER Working Paper No. 7552, available at: www.nber.org/papers/w7552 (accessed 30 November 2015).

Colquhoun, D. and Plested, A. (2014), "Why you should ignore altmetrics and other bibliometric nightmares", available at: www.dcscience.net/2014/01/16/why-you-shouldignore-altmetrics-and-other-bibliometric-nightmares (accessed 30 November 2015).

Costas, R., Zahedi, Z. and Wouters, P. (2015), "Do altmetrics correlate with citations?", *Extensive Comparison of Altmetric Indicators with Citations From a Multidisciplinary Perspective*, Vol. 66 No. 10, pp. 2003-2019.

Cronin, B., Snyder, H.W., Rosenbaum, H., Martinson, A. and Callahan, E. (1998), "Invoked on the web", *Journal of the American Society for Information Science*, Vol. 49 No. 14, pp. 1319-1328.

Dinsmore, A., Allen, L. and Dolby, K. (2014), "Alternative perspectives on impact: the potential of ALMs and altmetrics to inform funders about research impact", *PLOS Biology*, Vol. 12 No. 11, p. e1002003. doi: 10.1371/journal.pbio.1002003.

EPSRC (2011), "Evaluation of the PhD Plus pilot scheme", available at: www.epsrc.ac.uk/files/skills/evaluation-of-the-phd-plus-pilot-scheme-june-2011/ (accessed 30 November 2015).

ERA (2010), "What is a patent family name and how do I provide it?", available at: www.arc.gov.au/era/era_2010/archive/2010faq.htm (accessed 30 November 2015).

Eysenbach, G. (2011), "Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact", *Journal of Medical Internet Research*, Vol. 13 No. 4. doi: 10.2196/jmir.2012 (accessed 30 November 2015).

Fairclough, R. and Thelwall, M. (2015a), "More precise methods for national research citation impact comparisons", *Journal of Informetrics*, Vol. 9 No. 4, pp. 895-906.

Fairclough, R. and Thelwall, M. (2015b), "National research impact indicators from Mendeley readers", *Journal of Informetrics*, Vol. 9 No. 4, pp. 845-859.

Franceschet, M. and Costantini, A. (2011), "The first Italian research assessment exercise: a bibliometric perspective", *Journal of Informetrics*, Vol. 5 No. 2, pp. 275-291.

Hamilton, S. (2011), "Evaluation of the ESRC's participation in European collaborative research projects (ECRPs)", available at: www.esrc.ac.uk/_images/ECRP_full_report_tcm8-22049.pdf (accessed 30 November 2015).

Hammarfelt, B. (2014), "Using altmetrics for assessing research impact in the humanities", *Scientometrics*, Vol. 101 No. 2, pp. 1419-1430.

Haustein, S., Peters, I., Sugimoto, C.R., Thelwall, M. and Larivière, V. (2014), "Tweeting biomedicine: an analysis of tweets and citations in the biomedical literature", *Journal of the Association for Information Science and Technology*, Vol. 65 No. 4, pp. 656-669.

Jaffe, A.B. (2002), "Building programme evaluation into the design of public research-support programmes", *Oxford Review of Economic Policy*, Vol. 18 No. 1, pp. 22-34.

Kousha, K. and Thelwall, M. (2007), "Google Scholar citations and Google Web/URL citations: a multi-discipline exploratory analysis", *Journal of the American Society for Information Science and Technology*, Vol. 57 No. 6, pp. 1055-1065.

Kousha, K. and Thelwall, M. (2008), "Assessing the impact of disciplinary research on teaching: an automatic analysis of online syllabuses", *Journal of the American Society for Information Science and Technology*, Vol. 59 No. 13, pp. 2060-2069.

Kousha, K. and Thelwall, M. (2014), "An automatic method for extracting citations from Google Books", *Journal of the Association for Information Science and Technology*, Vol. 66 No. 2, pp. 309-320. doi: 10.1002/asi.23170.

Kousha, K. and Thelwall, M. (2015), "Web indicators for research evaluation, part 3: books and non-refereed outputs", *El Profesional de la Información*, Vol. 24 No. 6.

Kousha, K. and Thelwall, M. (in press), "An automatic method for assessing the teaching impact of books from online academic syllabi", *Journal of the Association for Information Science and Technology*, available at: www.scit.wlv.ac.uk/~cm1993/papers/SyllabiBookCitations.pdf (accessed 30 November 2015).

Li, X. and Thelwall, M. (2012), "F1000, Mendeley and traditional bibliometric indicators", *17th International Conference on Science and Technology Indicators, Vol. 3*, pp. 1-11.

Li, X., Thelwall, M. and Giustini, D. (2012), "Validating online reference managers for scholarly impact measurement", *Scientometrics*, Vol. 91 No. 2, pp. 461-471.

Maflahi, N. and Thelwall, M. (in press), "When are readership counts as useful as citation counts? Scopus vs Mendeley for LIS journals", *Journal of the Association for Information Science and Technology*.

Mas-Bleda, A., Thelwall, M., Kousha, K. and Aguillo, I.F. (2014), "Do highly cited researchers successfully use the social web?", *Scientometrics*, Vol. 101 No. 1, pp. 337-356.

Meagher, L. (2009), "Evaluation of the ESRC/MRC interdisciplinary studentship and postdoctoral fellowship scheme", available at: www.esrc.ac.uk/_images/Evaluation-of-ESRC-MRC-interdisciplinary-studentship-and-pdf-scheme_tcm8-24165.pdf (accessed 30 November 2015).

Meyer, M. (2003), "Academic patents as an indicator of useful research? A new approach to measure academic inventiveness", *Research Evaluation*, Vol. 12 No. 1, pp. 17-27.

Moed, H.F. (2006), *Citation Analysis in Research Evaluation*, Springer, Berlin.

Mohammadi, E. and Thelwall, M. (2013), "Assessing non-standard article impact using F1000 labels", *Scientometrics*, Vol. 97 No. 2, pp. 383-395.

Mohammadi, E., Thelwall, M. and Kousha, K. (in press), "Can Mendeley bookmarks reflect readership? A survey of user motivations", *Journal of the Association for Information Science and Technology*. doi: 10.1002/asi.23477.

Mohammadi, E., Thelwall, M., Haustein, S. and Larivière, V. (2015), "Who reads research articles? An altmetrics analysis of Mendeley user categories", *Journal of the Association for Information Science and Technology*, Vol. 66 No. 9, pp. 1832-1846. doi: 10.1002/asi.23286.

Mollett, A., Moran, D. and Dunleavy, P. (2011), "Using Twitter in university research, teaching and impact activities", London School of Economics and Political Science, London, available at: http://eprints.lse.ac.uk/38489/ (accessed 30 November 2015).

Narin, F. (1994), "Patent bibliometrics", *Scientometrics*, Vol. 30 No. 1, pp. 147-155.

Nederhof, A.J. and Van Raan, A.F. (1993), "A bibliometric analysis of six economics research groups: a comparison with peer review", *Research Policy*, Vol. 22 No. 4, pp. 353-368.

Neylon, C. and Wu, S. (2009), "Article-level metrics and the evolution of scientific impact", *Plos Biology*, Vol. 7 No. 11. doi: 10.1371/journal.pbio.1000242.

Oppenheim, C. (2000), "Do patent citations count?", in Cronin, B. (Ed.), *The Web of Knowledge*, Information Today, Inc., Medford, NJ, pp. 405-432.

Priem, J., Taraborelli, D., Groth, P. and Neylon, C. (2010), "Altmetrics: a manifesto", available at: http://altmetrics.org

REF2014 (2012), "Assessment framework and guidance on submissions", available at: www.ref.ac.uk/media/ref/content/pub/assessmentframeworkandguidanceonsubmissions/GOS%20including%20addendum.pdf

Rinia, E.J., Van Leeuwen, T.N., Van Vuren, H.G. and Van Raan, A.F. (1998), "Comparative analysis of a set of bibliometric indicators and central peer review criteria: evaluation of condensed matter physics in the Netherlands", *Research Policy*, Vol. 27 No. 1, pp. 95-107.

Shema, H., Bar-Ilan, J. and 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 American Society for Information Science and Technology*, Vol. 65 No. 5, pp. 1018-1027.

Shema, H., Bar-Ilan, J. and Thelwall, M. (2015), "How is research blogged? A content analysis approach", *Journal of the Association for Information Science and Technology*, Vol. 66 No. 6, pp. 1136-1149.

Shuai, X., Pepe, A. and Bollen, J. (2012), "How the scientific community reacts to newly submitted preprints: article downloads, Twitter mentions, and citations", *Plos One*, Vol. 7 No. 11.

Sud, P. and Thelwall, M. (2014), "Evaluating altmetrics", *Scientometrics*, Vol. 98 No. 2, pp. 1131-1143.

Taubes, G. (1993), *Bad Science: The Short Life and Weird Times of Cold Fusion*, Randonm House, New York, NY.

Thelwall, M. and Delgado, M. (2015), "Arts and humanities research evaluation: no metrics please, just data", *Journal of Documentation*, Vol. 71 No. 4, pp. 817-833.

Thelwall, M. and Fairclough, R. (2015), "The influence of time and discipline on the magnitude of correlations between citation counts and quality scores", *Journal of Informetrics*, Vol. 9 No. 3, pp. 529-541.

Thelwall, M., Haustein, S., Larivière, V. and Sugimoto, C. (2013), "Do altmetrics work? Twitter and ten other candidates", *Plos One*, Vol. 8 No. 5, p. e64841. doi: 10.1371/journal.pone.0064841.

Thelwall, M. and Kousha, K. (2008), "Online presentations as a source of scientific impact?: an analysis of powerpoint files citing academic journals", *Journal of the American Society for Information Science and Technology*, Vol. 59 No. 5, pp. 805-815.

Thelwall, M. and Kousha, K. (2015a), "Web indicators for research evaluation, part 1: citations and links to academic articles from the web", *El Profesional de la Información*, Vol. 24 No. 5, pp. 587-606.

Thelwall, M. and Kousha, K. (2015b), "Web indicators for research evaluation, part 2: social media metrics", *El Profesional de la Información*, Vol. 24 No. 5, pp. 607-620. doi: 10.3145/epi.2015.sep.09.

Thelwall, M. and Maflahi, N. (in press), "Guideline references and academic citations as evidence of the clinical value of health research", *Journal of the Association for Information Science and Technology*.

Thelwall, M. and Wilson, P. (2014), "Regression for citation data: an evaluation of different methods", *Journal of Informetrics*, Vol. 8 No. 4, pp. 963-971.

Thelwall, M. and Wilson, P. (in press), "Mendeley readership altmetrics for medical articles: an analysis of 45 fields", *Journal of the Association for Information Science and Technology*.

Thelwall, M., Tsou, A., Weingart, S., Holmberg, K. and Haustein, S. (2013), "Tweeting links to academic articles", *Cybermetrics*, Vol. 17 No. 1, available at: http://cybermetrics.cindoc.csic.es/articles/v17i1p1.html

Tijssen, R.J. (2001), "Global and domestic utilization of industrial relevant science: patent citation analysis of science-technology interactions and knowledge flows", *Research Policy*, Vol. 30 No. 1, pp. 35-54.

Torres-Salinas, D., Robinson-García, N., Jiménez-Contreras, E. and Delgado López-Cózar, E. (2012), "Towards a 'book publishers citation reports'. First approach using the 'Book Citation Index' ", *Revista Española de Documentación Científica*, Vol. 35 No. 4, pp. 615-620.

Vaughan, L. and Shaw, D. (2003), "Bibliographic and web citations: what is the difference?", *Journal of the American Society for Information Science and Technology*, Vol. 54 No. 14, pp. 1313-1322.

Waltman, L. and Costas, R. (2014), "F1000 recommendations as a potential new data source for research evaluation: a comparison with citations", *Journal of the Association for Information Science and Technology*, Vol. 65 No. 3, pp. 433-445.

Wilkinson, D., Sud, P. and Thelwall, M. (2014), "Substance without citation: evaluating the online impact of grey literature", *Scientometrics*, Vol. 98 No. 2, pp. 797-806.

Wouters, P. and Costas, R. (2012), "Users, narcissism and control – tracking the impact of scholarly publications in the 21st century", SURF foundation (report), available at: www.surf.nl/binaries/content/assets/surf/en/knowledgebase/2011/Users+narcissism+and+control.pdf (accessed 30 November 2015).

Zahedi, Z., Costas, R. and Wouters, P. (2014a), "How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications" *Scientometrics*, Vol. 101 No. 2, pp. 1491-1513

Zahedi, Z., Haustein, S. and Bowman, T. (2014b), "Exploring data quality and retrieval strategies for Mendeley reader counts", presentation at SIGMET Metrics 2014 Workshop, 5 November, available at: www.slideshare.net/StefanieHaustein/sigme-tworkshop-asist2014 (accessed 30 November 2015).

Zitt, M. (2012), "The journal impact factor: angel, devil, or scapegoat? A comment on JK Vanclay's article 2011", *Scientometrics*, Vol. 92 No. 2, pp. 485-503.

Zuccala, A. and Guns, R. (2013), "Comparing book citations in humanities journals to library holdings: scholarly use versus 'perceived cultural benefit'", *Proceedings of ISSI 2013 – 14th International Society of Scientometrics and Informetrics Conference, AIT Austrian Institute of Technology GmbH, Vienna*, pp. 353-360.

Corresponding author
Professor Mike Thelwall can be contacted at: M.Thelwall@wlv.ac.uk