JOURNAL OF ECONOMIC SURVEYS

doi: 10.1111/joes.12381

MEASURING SOCIETAL IMPACTS OF RESEARCH WITH ALTMETRICS? COMMON PROBLEMS AND MISTAKES

Mike Thelwall*

University of Wolverhampton

Abstract. The impact agenda in many countries has led to increased attempts to assess the societal impacts of research. Altmetrics, webometrics, and other alternative indicators have been proposed to support this task, and many journal articles have been written that exploit alternative indicators to investigate societal impacts. Nevertheless, methodological studies of many of these indicators have revealed that extreme care must be taken with gathering, aggregating, and interpreting them. This article gives an overview of current alternative indicators, summarizes empirical research, and reports a series of common problems and mistakes to avoid when using them. The main issues are: selecting indicators to match goals; aggregating them in a way sensitive to field and publication year differences; largely avoiding them in formal evaluations; understanding that they reflect a biased fraction of the activity of interest; and understanding the type of impact reflected rather than interpreting them at face value.

Keywords. Altmetrics; Kousha metrics; Societal impact; Webometrics

1. Introduction

The government-led drive in many countries to push academic research toward activities with societal impacts is known as the impact agenda (Gunn and Mintrom, 2016; Chubb *et al*., 2017). These impacts might include commercial, cultural, social, or health benefits and might apply to individual organizations or society as a whole; typically only impacts within academia are ignored. This has led to a culture in which academics may need to formally or informally evaluate the societal impact of their work, with quantitative evidence when possible. In the UK, for example, departments need to write a set of impact case studies for the Research Excellence Framework, which are a set of evidence-based narratives about how their research has led to societal benefits (Martin, 2011). The peer review scores on these are set to direct a quarter of UK block grant funding from 2022, with submissions that are not judged as being internationally excellent being likely to receive nothing. This internationally increasing importance of demonstrating societal impacts for research has created a new demand for alternative impact indicators. Citation counts from the Web of Science, Scopus, and Google Scholar (Martín-Martín *et al*., 2018) are not very useful for this because they only directly reflect impacts within the academic publishing system.

Measuring the societal impacts of research is difficult for multiple reasons. From the (incomplete) linear model of innovation (Godin, 2011), basic research sometimes leads to applied research and

*Corresponding author contact email: m.thelwall@wlv.ac.uk.

Data sharing

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Journal of Economic Surveys (2020) Vol. 0, No. 0, pp. 1–13

applications, and so the pathway to impact for basic research can be long. In addition, science is a complex network of interrelationships and so it is hard to quantify the contribution of any one piece of research. For example, while Google originated within academia and its outline has been described by academic papers (e.g., Brin and Page, 1998; Page *et al*., 1999), their references would not reveal most acacemic term contributions. Google's core innovation, PageRank, is motivated by citation analysis but the Google papers only cite one directly relevant primary article about citation analysis (a relatively obscure paper from 1971). While early attempts to measure nonacademic impacts focused on patents or patent citations as indicators of commercial impacts, the patchy use of patents within industry and other factors undermine their value (Meyer, 2000; Oppenheim, 2000). In addition, patents are largely irrelevant to noncommercial research benefits, such as for health, politics, or culture. In the last 20 years, new interest has developed in exploiting the web to generate alternative impact indicators for academic research, however.

2. Alternative Indicators: Altmetrics and Webometrics

The use of the web as the default host for public documents has led to an increasing range being available online. They can sometimes be harvested to generate alternative indicators for different types of impact. These have been called webometrics (Almind and Ingwersen, 1997). Altmetrics are a more recent development and are described later. Most webometrics require a multistage data collection exercise to harvest. The first stage is often a set of automated commercial search engine queries to identify relevant online citing documents. The following are some key examples.

- *Syllabus citations* count online academic syllabi that cite a given academic paper or book (Mas Bleda and Thelwall, 2018). This is an educational impact indicator (Kousha and Thelwall, 2008) that can be generated by automated search engine queries and filtering. For individual academics, the Open Syllabus Project website can be consulted for summary statistics (opensyllabus.org).
- *Patent citations* count online patents indexed by Google that cite a given academic paper. This is a commercial impact indicator (Orduna-Malea *et al*., 2017) that can be generated by automated search engine queries and filtering. More sophisticated querying and more comprehensive coverage can be achieved through the subscription-based services, such as the Derwent World Patents Index. Thus, the web version's advantage is reduced cost.
- *Wikipedia citations* count Wikipedia pages citing a given academic paper or book in the reference list or bibliography at the end of many pages. This is possibly a general knowledge transfer impact indicator (Kousha and Thelwall, 2017) that can be generated by automated search engine queries.
- *Gray literature citations* count pdf and/or word documents online that cite a given paper. This is possibly an organizational impact indicator (Wilkinson *et al*., 2014), unless the gray literature is taken from a narrow source (e.g., government reports). Relevant gray literature citations can be generated by automated search engine queries and filtering. In narrow contexts, a more specific interpretation might be placed on gray literature citations.
- *PowerPoint citations* count online PowerPoint presentations that cite a given academic paper. This may be primarily an academic and educational impact indicator (Thelwall and Kousha, 2008) and can be generated by automated search engine queries. In contrast, SlideShare presentation citations seem to reflect professional impact for the fields in which this site is important (Thelwall and Kousha, 2017b).
- *Blog citations* count (usually) scientific blogs translating research papers for a general audience (Shema *et al*., 2015). This is probably a public interest indicator (Shema *et al*., 2014) that can be generated by automated search engine queries and filtering, to a limited extent.
- *News citations* count news stories in specified news websites that cite a given academic journal (Kousha and Thelwall, 2019) or papers (Ortega, 2019). This is a public or media interest indicator

that can be generated by automated search engine queries and filtering. News citations can also be extracted from subscription-based sources, such as ProQuest's news portfolio.
- *Google Books citations* are citations from books indexed by Google Books. This is potentially an arts/humanities/educational/cultural interest indicator (Kousha and Thelwall, 2015) because of the importance of monographs and edited volumes for the arts and humanities. It can be generated by automated Google Books API queries and filtering. Citations from books are also available from Scopus and the Web of Science Book Citation Index, but these have more limited coverage (Kousha *et al*., 2011; Torres-Salinas *et al*., 2014).
- *Clinical guideline citations* count public clinical guidelines that cite a given academic paper. This is a health impact indicator (Thelwall and Maflahi, 2016) that can be generated by automated search engine queries and filtering. *Drug guide citations* (Thelwall *et al*., 2017) are similar.

A second generation of web-based alternative indicators has arisen from the social web. These count mentions or citations from social platforms, such as Twitter and Facebook, and are known as altmetrics (Priem *et al*., 2010). Like webometrics, they have been designed to give evidence of nonacademic impacts, but usually exploit more informal sources and are much easier to collect, leading to their use by a substantial minority of academics (Aung *et al*., 2019). With the partial exception of Mendeley (which can be used purely as a reference manager rather than an academic social network or reference sharing site), all altmetrics are derived from social environments. The remaining altmetrics reflect spaces that are open to, and could reasonably be used by, the general public (e.g., Twitter, Facebook, YouTube, and Reddit). These might, therefore, reflect aspects of general societal impact (but see below) rather than the typically more focused scope of webometrics.

In practice, the word altmetrics seems to be often used as a general term to encompass indicators from the social web as well as webometrics or even any impact indicator other than citation counts. The following are some key examples of social web altmetrics. Google+ has also been investigated but closed in April 2019 (see also: https://www.altmetric.com/about-our-data/our-sources/).

- *Tweets* count Twitter posts that reference an academic publication, typically by hyperlinking to a page for the article containing the article's DOI. This is possibly a public or academic interest or attention indicator (Thelwall *et al*., 2013; Costas *et al*, 2015; Mohammadi *et al*., 2018), depending on field, that can be generated by Twitter API queries.
- *Public Facebook wall posts* are similar to tweets (Thelwall *et al*., 2013; Costas *et al*, 2015; Mohammadi *et al*., 2020), but Facebook lacks a relevant API so it is only practical to gather mentions from a limited set of public Facebook pages rather than the whole site.
- *Mendeley readers* count the number of users of the Mendeley website that have registered an academic publication in their personal library. This is an academic and partly educational impact indicator (via reading: Mohammadi *et al*., 2016) that can be generated by Mendeley API queries (Zahedi *et al*., 2014).
- *YouTube* counts mentions of academic research in comments. These seem likely to be rare.
- *Reddit* counts citations in Reddit posts. These are rare (Thelwall *et al*., 2013) and from a specialist male-dominated audience (Duggan and Smith, 2013). Reddit is a news-based community discussion site.

In May 2020, there were two main commercial groups that systematically gather and sell alternative indicators of both types, Altmetric.com (Digital Science) and PlumX (Elsevier). These both sell evaluation analytics toolkits and data to universities and others so that they can exploit alternative indicators to evaluate their own societal impacts. Many of these indicators can be generated with free software, such as Webometric Analyst (lexiurl.wlv.ac.uk), or collected from the Crossref Event Data service (www.crossref.org/services/event-data). Researchers may also request free altmetric data from

Table 1. Problems or Mistakes Described in This Article.

Problem or mistake	Solution
Indicators not matching goals	Choose indicators to match project goals
Using sparse indicators (mostly 0) for small datasets	Use common indicators (e.g., Twitter and Facebook) or expand dataset size
Ignoring field differences	Compare scores only within fields or use a field normalized indicator
Ignoring publication year differences	Compare scores only within the same year (and field) or use a field normalized indicator
Calculating arithmetic means	Calculate geometric means or use log-normalized indicator
Reporting Pearson correlations	Report Spearman correlations
Overinterpreting correlations	Avoid making strong inferences from correlation coefficients
Treating indicators as unbiased	Report likely indicator biases
Uncritically interpreting indicator meaning	Critically evaluate the meaning of the indicator
Using hybrid indicators	Use separate indicators
Not reporting data collection date and method	Report data collection date and method
Ineffective literature review	Find and critically evaluate relevant altmetric literature
Use for formal evaluations	Avoid in formal evaluations, except when a special case can be made

Altmetric.com. The choice of source and processing methods can influence the results of a study (Ortega, 2018; Zahedi and Costas, 2018; Bar-Ilan *et al*., 2019), so care is needed when interpreting the results.

Alternative indicators of other types also exist, such as download counts from publishers (Shuai *et al*., 2012) or the Web of Science (Wang *et al*., 2016) and access statistics from academic social network sites like Academia.edu (Thelwall and Kousha, 2014) and ResearchGate (Yu *et al*., 2016; Thelwall and Kousha, 2017a), but these seem to be difficult to harvest for large systematic analyses.

3. Common Problems and Mistakes in Altmetrics Papers

This section provides a nonexhaustive list of analytical issues to be aware of when applying alternative indicators, recognizing the many challenges for effective use (Liu and Adie, 2013; Haustein, 2016). The list is drawn from my own experience as a referee of altmetrics papers and from previous literature reviews (Table 1). It does not include all issues, since it focuses on those that I have seen. These issues are sometimes mistakes that render the analysis invalid and are sometimes less-than-optimal methods that increase the chance of the results being misleading. The list excludes problems that are altmetric limitations rather than mistakes, such as the potential for manipulation of the scores, incomplete coverage of scholarly documents by altmetrics (e.g., for papers without DOIs or other standard identifiers for some altmetrics), and technical issues with definitions of academic fields. When relevant, solutions are described in detail.

Research articles exploiting alternative indicators tend to either evaluate the value of the indicators (by far the most common type of paper) or apply them to a task. Even research taking the latter approach needs to consider the value of the indicators used, however, and needs to understand how indicators are evaluated (Sud and Thelwall, 2014). The underlying reason for the number of mistakes in alternative

indicator papers may be that they have attracted wide interest among the scholarly community but their appropriate use requires thinking through many issues that might not be obvious at first.

3.1 *Indicators not Matching Goals*

A common error is to throw a basket of indicators, or a single hybrid altmetric, at a problem without a theoretical or problem-based reason for their selection. Given that altmetric data providers deliver a predefined package of alternative indicators and that there is a wide range that could potentially be used, it is important to construct a rationale for selecting the indicators to investigate. This rationale should be derived from the goals for their use. Nevertheless, the goal might be broad and exploratory in the sense of seeking evidence of any nonacademic impacts, in which all alternative indicators would be relevant, except perhaps Mendeley readers (primarily an academic impact indicator).

3.2 *Using Sparse Indicators for Small Datasets*

An occasional error is to use sparse indicators (i.e., with most articles having a score of 0) for small datasets, which is not useful. Except for Mendeley readers and tweets, most altmetric scores are likely to be 0 for most documents in any set to be evaluated, unless it is a preselected collection of high impact documents. The sparser the indicator, the greater the number of documents that need to be evaluated across to give useful information. This is due to two interrelated reasons. First, if there are few documents, then they might all score 0 for a sparse indicator, giving little information. Second, if there are a moderate number of documents such that a few are nonzero, then the confidence interval for the average is likely to be large, so that it is little use for differentiating between sets of documents. Thus, for example, Wikipedia citations have little value for comparing the impact of sets of documents unless they are preselected for high impact or there are thousands (e.g., Thelwall *et al*., 2016).

This issue can be resolved either by using less sparse indicators (e.g., Twitter and Mendeley), if relevant to the project goals, or expanding the dataset, if possible.

3.3 *Ignoring Field Differences*

Ignoring field differences is a common error. As with citation counts, average alternative indicator scores vary between fields. For example, tweet counts for cancer-related research are likely to be much higher than for pure (basic) mathematics research. Thus, it would not be fair to compare aggregate tweet counts between sets of documents that were not from the same field. Statistical analyses that combine articles from multiple fields are likely to give misleading results because of this. For example, a mixed set of maths and cancer research articles would generate a high correlation between tweet counts and citation counts because mathematics articles attract few tweets and citations, whereas cancer research attracts many tweets and citations. Thus, the high correlation would be caused by a spurious factor rather than any connection between tweets and citations. For this reason, a university-wide or Scopus-wide correlation or average would be unhelpful.

3.3.1 *Solution: Field Normalized Indicators or Percentiles*

The field differences issue can be solved in two different ways: calculating multiple averages, one for each field; or calculating a field normalized impact indicator that reports counts as the ratio to the average for the field and year of publication (Waltman *et al*., 2011). For example, both the Mean Normalized Citation Score and the Mean Normalized Log-transformed Citation Score (MNLCS) eliminate field differences by transforming all counts to a common scale, where 0 is the minimum, 1 is the world average, and higher scores are better (Thelwall, 2017). Ideally, a field normalized indicator like the MNLCS should be used

that is sensitive to skewing (e.g., through log transformations) for the most precise estimate of central tendency.

An alternative to field normalized indicators is to analyze within-field percentiles. For example, if article A is in the top 5% for Oncology but article B is in the top 1% for Victorian Studies, then the latter would have performed better within its field, irrespective of the original numerical values of the alternative indicators.

3.4 *Ignoring Publication Year Differences*

Ignoring publication year differences is another reasonably common error. Statistical analyses of multiyear datasets are likely to be misleading because the results could be due to age differences, unless field normalized indicators (which also normalize for age) or within-field, within year percentiles are used. This is because older articles are likely to have higher citation counts due to having had longer to be cited. Alternative indicator scores also naturally vary, on average, as a function of the publishing article age and, therefore, the analysis results of multiyear data would be partly or wholly due to time differences. To give an extreme example, since Twitter is a real-time communication platform, average tweet counts for papers published before Twitter started are likely to be substantially lower than average tweet counts for contemporary papers, so comparing them would be pointless.

3.5 *Calculating Arithmetic Means*

Inappropriately using arithmetic means is a very common error. Since altmetrics, webometrics, and citation counts are highly skewed, calculating the arithmetic mean of a set of scores does not give the best measure of central tendency. This is because the result may be dominated by a few high scores rather than reflecting typical values. The (offset) geometric mean is a better alternative. For this, add 1 to the scores, take the natural log, take the arithmetic mean of these log-transformed scores (i.e., apply the formula $\ln(1 + x)$), and transform back with $\exp(x) - 1$ (Thelwall and Fairclough, 2015). The MNLCS field normalized indicator uses the same log transformation to deal with skewed data. This also allows parametric confidence intervals to be calculated from the results.

3.6 *Reporting Pearson Correlations*

Correlations between alternative indicators and citation counts are commonly calculated as tests to validate the alternative indicators and show that they are nonrandom. Calculating Pearson correlations when they are inappropriate is an occasional error that I have made myself in the past. Correlation tests are sometimes used to investigate the relationship between alternative indicators and citation counts. Since altmetrics, webometrics, and citation counts are highly skewed (a small number of high scores), Pearson correlations are unhelpful unless the raw data are log-transformed first (or all converted to field normalized indicators) to reduce skewing. Pearson correlations are inappropriate because they are sensitive to outliers, which can greatly change the results, and skewing potentially causes outliers. If log transformation is not possible, then Spearman correlations should be used to compare indicators.

3.7 *Overinterpreting Correlations*

Correlation values are tricky to interpret because of discrete data issues (Thelwall, 2016), which means that the magnitude of a correlation is affected by the distribution of the data (e.g., the percentage nonzero) as well as the underlying relationship between the two variables correlated. In particular, when the mode is 0, it is technically unlikely to obtain a high correlation. This is a difficult factor to take into account

when using correlations, but researchers should be aware of this difficulty by at least not making strong inferences from correlation values.

3.8 *Treating Alternative Indicators as Unbiased*

Some studies seem to overlook or forget to mention that alternative indicators are biased, giving a misleading impression of the robustness of their results. All alternative indicators seem to reflect a very small fraction of impacts of the type that they are relevant to. For example, since between 1 in 12 and 1 in 20 publishing academics use Mendeley, according to one survey (Van Noorden, 2014), it cannot fully reflect academia. Similarly, if tweet counts were used as a societal impact indicator, then the biases would include national (it is rarely used in China), age (average usage differs between age groups), and types of impacts that people would not like to tweet about (e.g., treatments for embarrassing diseases).

Perhaps the best case in terms of impact type coverage is clinical guideline citations. The UK and other governments provide a wide-ranging set of these guidelines, including references to the supporting research, as selected by a panel of experts. These are a best case in the sense that medical professionals are supposed to consult these guidelines routinely in their practice. Medical guideline-based indicators are still partial for several reasons: not all conditions are covered by the guidelines; most countries do not publish these guidelines; the guidelines may cite meta-analyses rather than primary studies; and more recent research and papers that informed the papers cited in the guidelines are unrecognized by the indicator. Thus, research can have substantial impacts on clinical practice in one or more countries without attracting any guideline citations. Of course, the guidelines also do not cover other aspects of the health process, such as drug development, clinical methods innovation, and healthcare management.

At the other extreme, if tweets are used as an indicator of public interest in academic research, then it seems likely that a very small fraction of cases of public interest lead to tweets—perhaps less than one in ten thousand? This introduces a huge self-selection sampling bias in addition to making it highly likely that any given instance of public attention to academic research will have no reflection on Twitter. If the data are aggregated on a sufficiently large scale, then it might be reasonable to assume that the more tweeted body of research has generated more public interest because random factors tend to cancel out for large numbers. Nevertheless, sampling bias never cancels out, so it would be unfair to compare counts for sets of articles that may have different biases. For example, a set of papers from a department focusing on online communication is likely to have a much higher tweeted fraction of its papers attracting public interest than a similar department focusing on communication among the elderly.

Another aspect of bias is that the data collection process generating the alternative indicators is likely to introduce language and national biases. Examples include syllabus mentions (language-specific queries used to find them), Twitter and Facebook (not used in China), blogs (a limited set of blog sites can be searched), and news (the main news sites cannot easily be harvested for academic citations, and there are language and country biases from the main providers: Ortega, 2019). Thus, researchers should be careful not to overclaim from their results and to report likely biases to give the reader context.

3.9 *Uncritically Interpreting Alternative Indicators at Face Value*

An understandably common fallacy with altmetrics is to equate the type of impact that they *might* reflect with the type of impact that they *do* reflect. For example, since the vast majority of Twitter users are nonacademics, it would be reasonable to believe that tweet counts reflect societal attention or impact. This is flawed since it is logically possible that nonacademic tweeters never tweet about research so that tweet citations reflect only academic interest. In this case, one study suggests that a slight majority

of tweets about academic research are from nonacademics (Mohammadi *et al*., 2018), so tweets might reflect half academic, half nonacademic attention or impact, but there are likely to be field differences in the academic/public ratio.

3.9.1 *Solution: Attributing Meaning to Alternative Indicators*

Papers should give attention to attributing meaning to indicators. This is a major concern for general indicators (Twitter, Facebook, and to some extent PowerPoint citations, gray literature citations, Blog citations, Google Books citations, and Wikipedia) but not for narrow indicators with a clear role (e.g., syllabus mentions, clinical guideline citations, and Mendeley readers).

Most academic studies of alternative indicators have attempted to evaluate them by providing information about the meaning, if any, that could be attributed to them. There are multiple reasonable evaluation methods (Sud and Thelwall, 2014), but the most common technique is to correlate mature alternative indicator values with citation counts to check if there is an association. On the basis that low-quality work will attract little interest of any type, and that research generating substantial impact of any type will tend to attract follow-up research that will cite it, it is reasonable to expect any genuine impact indicator to correlate positively with citations. A zero correlation would instead suggest that the indicator does not reflect impact but is either random or reflects something irrelevant (e.g., the amusingness of the title). The prevalence of bots on Twitter (Haustein *et al*., 2016; Didegah *et al*., 2018) and users tending to tweet shorter articles and editorials (Haustein *et al*., 2015) illustrate the importance of correlation tests in checking whether the overall results are meaningful. Correlation tests are, therefore, important even though the goal of alternative indicator use is not to reflect citation impact.

For alternative indicators for which the type of impact is not clear, content analysis of citation sources, surveys, and interviews with creators can be used to help decide what they reflect. For example, content analyses of tweets suggest that they reflect interest in articles rather than endorsement, use, or engagement with research (Thelwall *et al*., 2013; Holmberg, and Thelwall, 2014; Robinson-Garcia *et al*., 2017). In contrast, a survey of Mendeley users suggests that they mostly add articles to their libraries when they have read them or intend to read them (Mohammadi *et al*., 2016). Thus, it is reasonable to interpret Mendeley "reader counts" as essentially counts of (Mendeley-using) readers.

Even after empirical evaluations, the type of impact reflected by an alternative indicator, if any, may not be clear (e.g., Twitter, Facebook, and Wikipedia). In such cases, analyses should be explicit about this lack of clarity when reporting the scores.

3.10 *Using Hybrid Indicators*

Some studies of altmetrics have used hybrid indicators, such as the Altmetric Attention Score (aggregating multiple altmetrics into a single score, using a weighted sum: www.altmetric.com/blog/the-altmetric-score-is-now-the-altmetric-attention-score/) or a variant of the h-index (compounding impact and publication quantity). It is almost always preferable to analyze indicators separately because they have different interpretations and therefore a hybrid indicator, such as the Altmetric Attention Score, cannot be given a robust interpretation. This is compounded by the (understandable) lack of empirical support for the weightings used in the Altmetric Attention Score. This score is very useful as a quick statistic on publisher websites, where the reader can click on the Altmetric badge for a breakdown of a score, but it should not be used in research applications.

The hybridity issue also applies to the h-index: impact and quantity should be analyzed separately to give finer-grained information. For example, the h-index is biased against women, who are more likely to take career breaks. If publication counts and average impact are analyzed separately, then the one that is not biased against women (average citations) can be used for a fairer evaluation.

3.11 *Not Reporting the Data Collection Date and Method*

This is an occasional accidental omission. The exact date of data collection is essential because alternative indicators vary over time and so the time window between publication and data collection affects the results of a study. While this has little impact for fast accumulating altmetrics, such as Tweeter counts or Facebook Wall posts, it affects the slower accumulating altmetrics, such as Mendeley reader counts, and slow webometrics, such as policy document citations (Fang and Costas, 2020). The reader cannot properly interpret the results or compare them to related studies without relevant date information.

3.12 *Failing to Conduct an Effective Literature Review*

Medical altmetric studies seem to often ignore prior altmetric research, presumably because their literature search uses Medline, which indexes few altmetrics articles. Scopus, the Web of Science, Google Scholar, and Dimensions are all better choices for altmetrics. An effective literature search would pick up all the issues mentioned in the current paper, ensuring that the study was well designed and situated its results in the context of relevant prior literature. It could also identify general methods advice articles (Sud and Thelwall, 2014) to help with this.

3.13 *Using Alternative Indicators in Formal Evaluations*

This is a potentially disastrous error relating to practical applications of altmetrics. Nearly all alternative indicators lack quality control and are relatively easy to manipulate accidentally or deliberately. The main exceptions are clinical guidelines, drug guidelines, and gray literature citations from quality-controlled collections. The weaker exceptions (manipulation is possible but may not be straightforward) are Google Books citations, news citations, and syllabus mentions. For example, it would be easy to generate a series of temporary email accounts and use them to create dummy Twitter and Mendeley accounts. The Twitter accounts could then tweet links to any desired research papers, increasing their tweet counts, and the Mendeley accounts could be loaded with libraries of selected academic papers, increasing their Mendeley reader counts. If any kind of indicator is used in formal evaluations, then care must be taken to ensure that they are used responsibly so that those evaluated are not disadvantaged and they do not have negative unintended consequences (Wilsdon *et al*., 2015).

For research evaluations where those evaluated are told in advance, only the main exceptions should normally be allowed. This is because academics have shown willingness to game any susceptible indicators (Zimmerman, 2013). Weaker indicators could only be used if it is judged that the likelihood of identifying manipulation is greater than the gain from the higher scores. This is effectively the position in the UK REF impact case studies, where academics may include alternative indicators in support of a narrative case for impact (www.ref.ac.uk). An alternative would be having a strong honesty statement for the evaluated researchers to raise the cost of deliberate manipulation. A case could also be made for this exception to apply to researchers adding alternative indicators to their CVs since they directly take ownership of the data and it can be evaluated in context with the other information in the document (Piwowar and Priem, 2013).

For some research evaluations, the evaluation team decides after receiving a submission how to evaluate it (e.g., Belgium). In this case, the use of alternative indicators is reasonable if the evaluators judge that it is unlikely for those being evaluated to have guessed that a set of alternative indicators might be used and manipulate them. This also applies to research evaluations where the outcomes do not have negative consequences for the researchers or other key stakeholders. For example, research funder

evaluations used to monitor the effectiveness of funding streams (Dinsmore *et al*., 2014; Thelwall *et al*., 2016) may be purely formative.

Altmetrics can presumably be used for formative self-evaluations without fear of deliberate manipulation (Wouters and Costas, 2012) and this seems to be their most common current use.

4. Conclusions and Recommendations

Alternative indicators, including social media-based altmetrics and webometrics, provide the potential to reflect different types of impacts to that of citation counts, or to give earlier impact evidence. This can help with formative evaluations (but see above) and self-evaluations (Wouters and Costas, 2012), as well as for studying science itself. Nevertheless, social media altmetrics in particular must be used cautiously, without assuming what kind of impact they reflect. For all alternative indicators, it is also important to derive aggregate indicators carefully and recognize that they only reflect a partial and potentially biased subset of the impact type that they reflect. For researchers using alternative indicators to assess the impacts of sets of publications or knowledge flows, it is important to be clear in the results that the data present one perspective rather than definitive picture. In summary, the key stages are (with the problematic areas discussed above in italics):

1. Conduct a literature search and critical analysis of altmetrics literature to understand the *potentials and limitations of altmetrics and appropriate methods* to use them.
2. Formulate the research design, including the scope (publications to be examined), the individual altmetrics to be used (*matching the goals*), and the analysis methods.
3. Collect the necessary alternative indicator data, *recording the method and access dates*.
4. Analyze the data using *appropriate statistical techniques*.
5. Critically evaluate the results, being careful *not to overclaim* from the results and *not to make assumptions about the meaning* of any altmetrics.

Despite the many limitations of altmetrics, they offer, for the first time, a relatively straightforward way to gather data about some of the many different societal impacts of research. Thus, they are likely to continue to be part of the toolkits of research evaluators and academics studying science for the foreseeable future.

References

Almind, T.C. and Ingwersen, P. (1997) Informetric analyses on the World Wide Web: methodological approaches to 'webometrics'. *Journal of Documentation* 53(4): 404–426.

Aung, H.H., Zheng, H., Erdt, M., Aw, A.S., Sin, S.C.J. and Theng, Y.L. (2019) Investigating familiarity and usage of traditional metrics and altmetrics. *Journal of the Association for Information Science and Technology* 70(8): 872–887.

Bar-Ilan, J., Halevi, G. and Milojević, S. (2019) Differences between altmetric data sources — a case study. *Journal of Altmetrics* 2(1): 1.

Brin, S. and Page, L. (1998) The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30(1–7): 107–117.

Chubb, J., Watermeyer, R. and Wakeling, P. (2017) Fear and loathing in the academy? The role of emotion in response to an impact agenda in the UK and Australia. *Higher Education Research & Development* 36(3): 555–568.

Costas, R., Zahedi, Z. and Wouters, P. (2015) Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology* 66(10): 2003–2019.

Didegah, F., Mejlgaard, N. and Sørensen, M.P. (2018) Investigating the quality of interactions and public engagement around scientific papers on Twitter. *Journal of Informetrics* 12(3): 960–971.

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* 12(11): e1002003.
Duggan, M. and Smith, A. (2013) 6% of online adults are reddit users. Pew Internet & American Life Project. https://www.pewresearch.org/internet/2013/07/03/6-of-online-adults-are-reddit-users/ (Accessed 20 June 2020).
Fang, Z. and Costas, R. (2020) Studying the accumulation velocity of altmetric data tracked by Altmetric. com. *Scientometrics* 123(3): 1077–1101.
Godin, B. (2011) The linear model of innovation: Maurice Holland and the research cycle. *Social Science Information* 50(3–4): 569–581.
Gunn, A. and Mintrom, M. (2016) Higher education policy change in Europe: academic research funding and the impact agenda. *European Education* 48(4): 241–257.
Haustein, S., Bowman, T.D., Holmberg, K., Tsou, A., Sugimoto, C.R. and Larivière, V. (2016) Tweets as impact indicators: examining the implications of automated "bot" accounts on Twitter. *Journal of the Association for Information Science and Technology* 67(1): 232–238.
Haustein, S., Costas, R. and Larivière, V. (2015) Characterizing social media metrics of scholarly papers: the effect of document properties and collaboration patterns. *PLoS One* 10(3): e0120495.
Haustein, S. (2016) Grand challenges in altmetrics: heterogeneity, data quality and dependencies. *Scientometrics* 108(1): 413–423.
Holmberg, K. and Thelwall, M. (2014) Disciplinary differences in Twitter scholarly communication. *Scientometrics* 101(2): 1027–1042.
Kousha, K., Thelwall, M. and Rezaie, S. (2011) Assessing the citation impact of books: the role of Google Books, Google Scholar, and Scopus. *Journal of the American Society for Information Science and Technology* 62(11): 2147–2164.
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*, 59(13): 2060–2069.
Kousha, K. and Thelwall, M. (2015) An automatic method for extracting citations from Google Books. *Journal of the American Society for Information Science and Technology* 66(2): 309–320.
Kousha, K. and Thelwall, M. (2017) Are Wikipedia citations important evidence of the impact of scholarly articles and books? *Journal of the Association for Information Science and Technology* 68(3): 762–779.
Kousha, K. and Thelwall, M. (2019) An automatic method to identify citations to journals in news stories: a case study of the UK newspapers citing Web of Science journals. *Journal of Data and Information Science* 4(3): 73–95.
Liu, J. and Adie, E. (2013) Five challenges in altmetrics: a toolmaker's perspective. *Bulletin of the American Society for Information Science and Technology* 39(4): 31–34.
Martín-Martín, A., Orduna-Malea, E., Thelwall, M. and López-Cózar, E.D. (2018) Google Scholar, Web of Science, and Scopus: a systematic comparison of citations in 252 subject categories. *Journal of Informetrics* 12(4): 1160–1177.
Martin, B.R. (2011) The research excellence framework and the 'impact agenda': are we creating a Frankenstein monster? *Research Evaluation* 20(3): 247–254.
Mas Bleda, A. and Thelwall, M. (2018) Assessing the teaching value of non-English academic books: the case of Spain. *Revista Española de Documentación Científica* 41(4): e222.
Meyer, M. (2000) What is special about patent citations? Differences between scientific and patent citations. *Scientometrics* 49(1): 93–123.
Mohammadi, E., Barahmand, N. and Thelwall, M. (2020) *Who Shares Health and Medical Scholarly Articles on Facebook?* Learned Publishing.
Mohammadi, E., Thelwall, M. and Kousha, K. (2016) Can Mendeley bookmarks reflect readership? A survey of user motivations. *Journal of the Association for Information Science and Technology* 67(5): 1198–1209.
Mohammadi, E., Thelwall, M., Kwasny, M. and Holmes, K. (2018) Academic information on Twitter: a user survey. *PLoS One* 13(5): e0197265.
Oppenheim, C. (2000) Do patent citations count. In H.B. Atkins & B. Cronin (eds.), *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield* (pp. 405–432). Medford, NJ: Information Today Inc.

Orduna-Malea, E., Thelwall, M. and Kousha, K. (2017) Web citations in patents: evidence of technological impact? *Journal of the Association for Information Science and Technology* 68(8): 1967–1974.
Ortega, J.L. (2018) Reliability and accuracy of altmetric providers: a comparison among Altmetric.com, PlumX and Crossref Event Data. *Scientometrics* 116(3): 2123–2138.
Ortega, J.L. (2019) Availability and audit of links in altmetric data providers: link checking of blogs and news in Altmetric.com, Crossref Event Data and PlumX. *Journal of Altmetrics* 2(1) paper 4. https://doi.org/10.29024/joa.14.
Page, L., Brin, S., Motwani, R. and Winograd, T. (1999) The PageRank citation ranking: bringing order to the web. Stanford InfoLab.
Piwowar, H. and Priem, J. (2013) The power of altmetrics on a CV. *Bulletin of the American Society for Information Science and Technology* 39(4): 10–13.
Priem, J., Taraborelli, D., Groth, P. and Neylon, C. (2010) Altmetrics: a manifesto. http://altmetrics.org/manifesto/ (Accessed 20 June 2020).
Robinson-Garcia, N., Costas, R., Isett, K., Melkers, J. and Hicks, D. (2017) The unbearable emptiness of tweeting—about journal articles. *PLoS One* 12(8): e0183551.
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 Association for Information Science and Technology* 65(5): 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* 66(6): 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* 7(11): e47523.
Sud, P. and Thelwall, M. (2014) Evaluating altmetrics. *Scientometrics* 98(2): 1131–1143.
Thelwall, M. and Fairclough, R. (2015) Geometric journal impact factors correcting for individual highly cited articles. *Journal of Informetrics* 9(2): 263–272.
Thelwall, M., Haustein, S., Larivière, V. and Sugimoto, C. (2013) Do altmetrics work? Twitter and ten other candidates. *PLoS One* 8(5): e64841.
Thelwall, M., Kousha, K. and Abdoli, M. (2017) Is medical research informing professional practice more highly cited? Evidence from AHFS DI Essentials in Drugs.com. *Scientometrics* 112(1): 509–527.
Thelwall, M., Kousha, K., Dinsmore, A. and Dolby, K. (2016) Alternative metric indicators for funding scheme evaluations. *Aslib Journal of Information Management* 68(1): 2–18.
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* 59(5): 805–815.
Thelwall, M. and Kousha, K. (2014) Academia.edu: social network or academic network? *Journal of the Association for Information Science and Technology* 65(4): 721–731.
Thelwall, M. and Kousha, K. (2017a) ResearchGate articles: age, discipline, audience size, and impact. *Journal of the Association for Information Science and Technology* 68(2): 468–479.
Thelwall, M. and Kousha, K. (2017b) SlideShare presentations, citations, users and trends: a professional site with academic and educational uses. *Journal of the Association for Information Science and Technology* 68(8): 1989–2003.
Thelwall, M. and Maflahi, N. (2016) Guideline references and academic citations as evidence of the clinical value of health research. *Journal of the Association for Information Science and Technology* 67(4): 960–966.
Thelwall, M., Tsou, A., Weingart, S., Holmberg, K. and Haustein, S. (2013) Tweeting links to academic articles. *Cybermetrics: International Journal of Scientometrics, Informetrics and Bibliometrics* 17: 1–8.
Thelwall, M. (2016) Interpreting correlations between citation counts and other indicators. *Scientometrics* 108(1): 337–347.
Thelwall, M. (2017) Three practical field normalised alternative indicator formulae for research evaluation. *Journal of Informetrics* 11(1): 128–151.
Torres-Salinas, D., Robinson-Garcia, N., Campanario, J.M. and López-Cózar, E.D. (2014) Coverage, field specialisation and the impact of scientific publishers indexed in the Book Citation Index. *Online Information Review* 38(1): 24–42.

Van Noorden, R. (2014) Online collaboration: scientists and the social network. *Nature* 512(7513): 126.
Waltman, L., van Eck, N.J., van Leeuwen, T.N., Visser, M.S. and van Raan, A.F. (2011) Towards a new crown indicator: an empirical analysis. *Scientometrics* 87(3): 467–481.
Wang, X., Fang, Z. and Sun, X. (2016) Usage patterns of scholarly articles on Web of Science: a study on Web of Science usage count. *Scientometrics* 109(2): 917–926.
Wouters, P. and Costas, R. (2012) Users, narcissism and control: tracking the impact of scholarly publications in the 21st century. Utrecht, The Netherlands: SURFfoundation.
Wilkinson, D., Sud, P. and Thelwall, M. (2014) Substance without citation: evaluating the online impact of grey literature. *Scientometrics* 98(2): 797–806.
Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S. and Tinkler, J. (2015) The metric tide. The metric tide: independent review of the role of metrics in research assessment and management. https://responsiblemetrics.org/the-metric-tide/ (Accessed 20 June 2020).
Yu, M.C., Wu, Y.C.J., Alhalabi, W., Kao, H.Y. and Wu, W.H. (2016) ResearchGate: an effective altmetric indicator for active researchers? *Computers in Human Behavior* 55: 1001–1006.
Zahedi, Z. and Costas, R. (2018) General discussion of data quality challenges in social media metrics: extensive comparison of four major altmetric data aggregators. *PLoS One* 13(5): e0197326.
Zahedi, Z., Haustein, S. and Bowman, T. (2014) Exploring data quality and retrieval strategies for Mendeley reader counts. In SIG/MET Workshop, ASIS&T 2014 Annual Meeting, Seattle. www.asis.org/SIG/SIGMET/data/uploads/sigmet2014/zahedi.pdf (Accessed 20 June 2020).
Zimmermann, C. (2013) Academic rankings with RePEc. *Econometrics* 1(3): 249–280.