STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

Convergent validity of altmetrics and case studies for assessing societal impact: an analysis based on UK Research Excellence Framework (REF) data[1]

Lutz Bornmann*, Robin Haunschild** and Jonathan Adams***

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

** *R.Haunschild@fkf.mpg.de*
Max Planck Institute for Solid State Research, Heisenbergstr. 1, 70569 Stuttgart, Germany

*** *jonathan.adams@clarivate.com*
The Policy Institute at King's, King's College London, 22 Kingsway, London WC2B 6LE, UK, and Institute for Scientific Information, Clarivate Analytics

Introduction

The significance of research publications has historically focused on their impact within science as measured by citations. Recently, however, there has been a broadening of the impact concept stimulated by governmental interest about the returns from research to society and the economy. This has led to the appearance of various definitions of societal impact. We highlight one definition, which was formulated as part of an overview of the societal impact literature and is very broad in its nature: "Research has a societal impact when auditable or recorded influence is achieved upon non-academic organisation(s) or actor(s) in a sector outside the university sector itself – for instance, by being used by one or more business corporations, government bodies, civil society organisations, media or specialist/professional media organisations or in public debate" (Wilsdon et al., 2015, p. 6). According to the National Research Council (2014) "no high-quality metrics for measuring societal impact currently exist that are adequate for evaluating the impacts of federally funded research on a national scale" (p. 70). One possible candidate that has been proposed for quantitatively indexing societal impact is altmetrics data, but no systematic evidence for or against this suggestion yet exists (Haustein, Bowman, et al., 2014).

Following earlier studies by Bornmann (2014), Bornmann (2015b), and Ravenscroft, Liakata, Clare, and Duma (2017) we address the question of the convergent and discriminant validity of altmetrics data for measuring societal impact. Using data from the UK Research Excellence Framework (REF) and the company Altmetric (see www.altmetric.com), we investigate in this study whether societal impact can be measured by using altmetrics or not (we use the term 'measure' as this is common practice but we note that indexed data are in fact indicators

[1] The bibliometric data used in this paper are from a custom database of the Competence Center for Bibliometrics (http://www.bibliometrie.info/). Altmetrics data were used from a locally maintained database with data shared with us by the company Altmetric on October 02, 2017. The REF output data were downloaded from http://results.ref.ac.uk/DownloadSubmissions/ByForm/REF2 on September 28, 2017. The REF case study IDs and corresponding DOIs were shared with us by Digital Science on December 11, 2017.

and there can be no direct impact measurement). We compare the impact of two groups of publications:
(1) Publications referenced as underpinning research in impact Case Studies (PCS): REF impact case studies are short (four page) documents each containing up to six relevant references and used by UK universities to describe the socio-economic impact of their research (Derrick, Meijer, & van Wijk, 2014). We predict high societal, but variable citation impact for these publications.
(2) Publications submitted as REF Research Outputs (PRO): To evidence academic achievement, UK institutions submit four research publications for each selected research staff member. We predict variable and usually low societal, but high citation impact for these publications.
To determine convergent validity, our purpose is to study whether two different approaches to the assessment of societal impact (case studies and altmetrics) are able to index the same construct in a similar way (Picardi & Masick, 2013; Thorngate, Dawes, & Foddy, 2009). We further study discriminant validity by analyzing the comparative societal and citation impact of PCS and PRO. We used the MHq' indicator proposed by Bornmann and Haunschild (in press) to analyze convergent and discriminant validity, because the indicator has been developed (and successfully tested) as a field-normalized indicator for count data with many zeros, e. g., altmetrics data.

Methods

Description of altmetrics
Altmetrics cover a diverse range of data (e.g., views, downloads, clicks, notes, saves, tweets, shares, likes, recommends, tags, posts, trackbacks, discussions, bookmarks, and comments). In this study, we have included six altmetrics that are frequently investigated in altmetrics' studies. A detailed overview of research on these altmetrics can be found in Thelwall (2017) and Sugimoto, Work, Larivière, and Haustein (2017).
Blogs which are online narratives are one of the earliest social media platforms (Bik & Goldstein, 2013). These blogs are also written about scholarly papers, which are cited in a formal or informal way (Bar-Ilan, Shema, & Thelwall, 2014). The citations can be counted.
Facebook is a widely used social networking platform. Since users on Facebook may share information on papers with other users, mentions of papers in posts can be counted (Ringelhan, Wollersheim, & Welpe, 2015).
News attention (e.g., by the New York Times) refers to scientific papers mentioned (via direct links or unique identifiers) in news reports (Priem, 2014). Thus, the attention can be counted.
Mentions of papers in **policy-related documents** are now analyzed for altmetrics, although this is a recent innovation. These mentions are discovered by text mining solutions in corresponding databases from governments (e.g. UK Ministry document archives) and intergovernmental organizations (e.g. World Health Organization) (Arthur, 2016; Bornmann, Haunschild, & Marx, 2016; Haunschild & Bornmann, 2017; Liu, 2014).
Twitter is a popular microblogging platform. Tweets of up to 140 (now 280) characters from users to followers can contain DOI references trackable to scientific papers (Haustein, Peters, Sugimoto, Thelwall, & Larivière, 2014).
Wikipedia is a free encyclopedia platform with editable content (Mas-Bleda & Thelwall, 2016). Contributors include references to scholarly papers (Serrano-López, Ingwersen, & Sanz-Casado, 2017).

Mantel-Haenszel quotient (MHq')

Bornmann and Haunschild (in press) proposed the use of the MHq' indicator as a field- and time-normalized altmetrics indicator, because the indicator is especially designed for count data with many zeros. Occurrence of many zeros has been observed in most altmetrics data which means that the usual normalization procedures in bibliometrics should not be applied to these altmetrics. The following explanation of the MHq' indicator is based on Bornmann and Haunschild (in press).

In contrast to many other normalized indicators in bibliometrics, MHq' is not calculated on the single paper level, but on an aggregated level considering field and time of publication. For the impact comparison of publication sets (here PRO and PCS) with reference sets, the 2×2 cross tables (which are pooled) consist of the number of papers mentioned and not mentioned in subject category and publication year combinations *f*. Thus, in the 2×2 subject-specific cross table with the cells a_f, b_f, c_f, and d_f (see Table 1), a_f is the number of mentioned papers in set *g* in subject category and publication year *f*, b_f is the number of not mentioned papers in set *g* in subject category and publication year *f*, c_f is the number of mentioned papers in subject category and publication year *f*, d_f is the number of not mentioned papers published in subject category and publication year *f*. As MHq' compares groups of papers, the papers of set *g* are not part of the papers in the world.

Table 1. 2 x 2 subject-specific cross table.

	Number of mentioned papers	**Number of not mentioned papers**
Group *g*	a_f	b_f
World	$c_f' = c_f - a_f$	$d_f' = d_f - b_f$

We start by defining some dummy variables for the MH analysis:

$$R_f = \frac{a_f d_f'}{n_f} \text{ and } R = \sum_{f=1}^{F} R_f, \quad (1)$$

$$S_f = \frac{b_f c_f'}{n_f} \text{ and } S = \sum_{f=1}^{F} S_f, \quad (2)$$

$$P_f = \frac{a_f + d_f'}{n_f} \text{ and } Q_f = 1 - P_f \quad (3)$$

Where $n_f = a_f + b_f + c_f' + d_f'$

MHq' is simply:

$$\mathrm{MHq}' = \frac{R}{S} \quad (4)$$

The CIs for MHq' are calculated following Fleiss, Levin, and Paik (2003). The variance of ln MHq' is estimated by:

$$\widehat{Var}(\ln MHq') = \frac{1}{2}\left\{\frac{\sum_{f=1}^{F} P_f R_f}{R^2} + \frac{\sum_{f=1}^{F}(P_f S_f + Q_f R_f)}{RS} + \frac{\sum_{f=1}^{F} Q_f S_f}{S^2}\right\} \quad (5)$$

The confidence interval for the MHq' can be constructed with

$$MHq_L' = \exp\left[\ln(MHq') - 1.96\sqrt{\widehat{Var}[\ln(MHq')]}\right] \quad (6)$$

$$MHq_U' = \exp\left[\ln(MHq') + 1.96\sqrt{\widehat{Var}[\ln(MHq')]}\right] \quad (7)$$

Dataset used

The REF output data including publication DOIs where available (outputs include articles, books, proceedings and audio and visual material) were downloaded from http://results.ref.ac.uk/DownloadSubmissions/ByForm/REF2 on September 28, 2017. Some 149,616 of 250,043 publications (59.8%) submitted as REF output papers had a DOI. The REF case study IDs and their corresponding DOIs were shared with us by Digital Science (see https://www.digital-science.com) on December 11, 2017 (Digital Science, 2016). Of the papers referenced in case studies, 25,313 had a DOI. We used the DOIs as a unique identifier to add citation and altmetric counts as metadata for each publication record.

Citation data from Elsevier's Scopus database (see https://www.scopus.com) were used via a custom database of the Competence Center for Bibliometrics (see http://www.bibliometrie.info/) which was last updated on April 2017. All papers submitted as PRO or PCS were matched via their DOI with the Scopus database. The number of citations and the Scopus subject area (see https://service.elsevier.com/app/answers/detail/a_id/15181/supporthub/scopus/) were appended to each DOI. For the papers referenced in case studies (PCS), 17,525 (69.2%) could be found in the Scopus database via their DOI. For the papers submitted as REF outputs (PRO), 126,694 (84.7%) could be matched via their DOI to the Scopus database.

Citations were determined using a two-year citation window for all papers published before 2015. The two-year citation window is shorter than is typical in bibliometrics studies but was appropriate for these relatively recent publications. Furthermore, it is a compromise between sufficient time to register impact and the shortened time to be used with the MHq' especially designed for count data with many zeros. Even so, the papers published in 2015 and 2016 (n=49 papers) were not included in the analysis because the citation window would be too short. The Scopus subject areas were aggregated to a higher level, i.e., subject codes ABCD with C,D>0 were merged into the subject code AB00. Some papers were assigned to multiple aggregated Scopus subject areas. We constructed overlapping Scopus subject areas from this multiple classification, which in the following are referred to as fields (Rons, 2012, 2014). In total, we obtained 732 fields.

Some papers were mentioned in multiple case studies and some papers were submitted multiple times in the output of the REF by different Units of Assessment (UOAs, which conform approximately to a field or discipline) and by individual academics from different institutions within the same UOA. Therefore, some duplicated papers are contained in our dataset. In total, 138,309 papers (136,793 papers with unique DOIs) are included in our analysis.

Altmetrics data were used from a locally maintained database with data shared with us by the company Altmetric on 02 October 2017. We appended a mention count to each DOI using six altmetrics (see section 0). A DOI not known to the altmetrics database was counted as 'not mentioned'.

Results

Table 2 shows the different comparisons in this study and expected outcomes. We compare altmetrics scores (e.g. tweets) with traditional citation scores. We expect higher altmetrics scores for PCS than for PRO and higher citation scores for PRO than for PCS. The results of Digital Science (2016) show, however, that a large part of the PCS were also PRO in the previous REFs. We therefore aggregated the data into three groups:

(1) PCS (not part of PRO): 11,822 papers

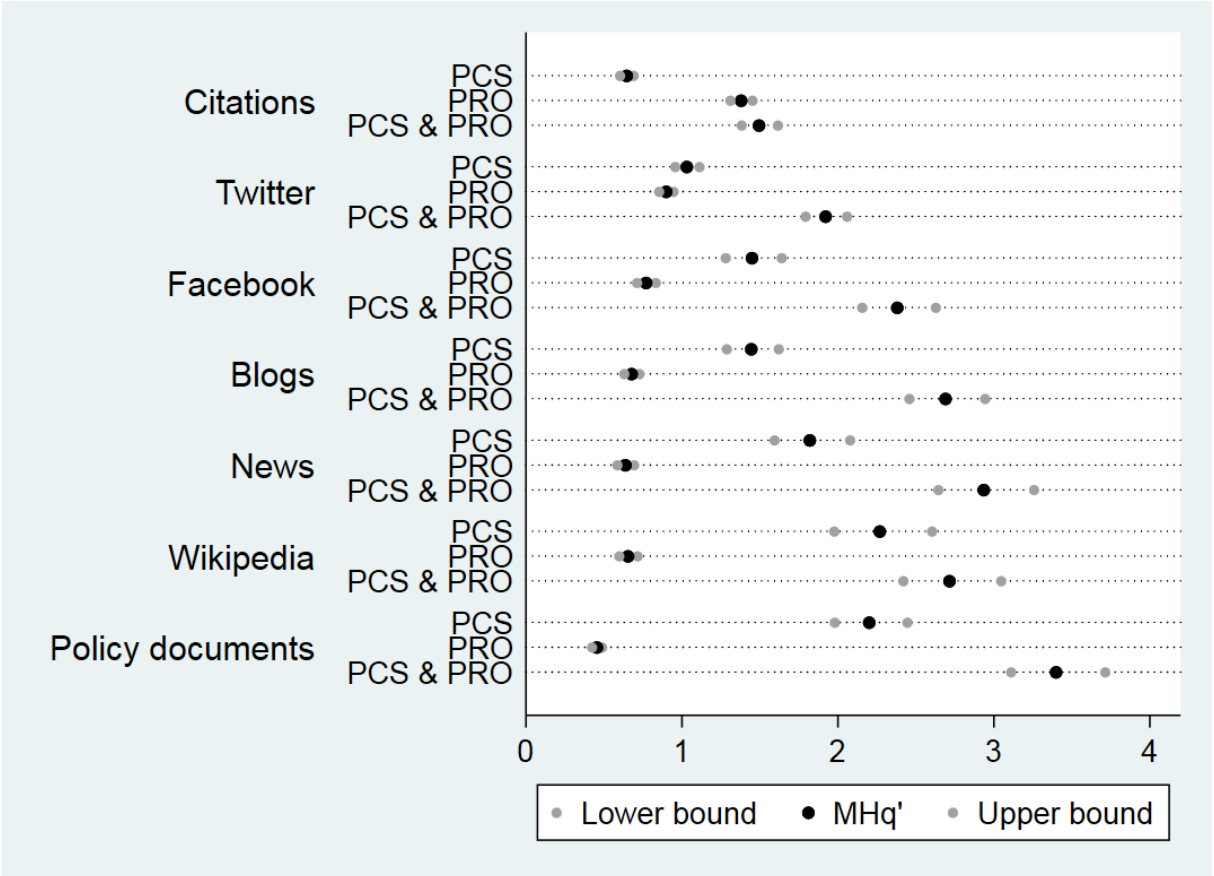
(2) PRO (not part of PCS): 120,784 papers
(3) PCS & PRO (PRO, part of PCS): 5,703 papers
With the separation into three groups, we identify two groups (1) and (2) that do not overlap in terms of publications. The expected metrics scores for all groups are shown in Table 2. We might reasonably expect the highest scores for the [PCS & PRO] group because these papers should attract attention in multiple impact dimensions.

Table 2. Analyzing convergent and discriminant validity in this study: expected metrics score.

	PCS (not part of PRO)	**PRO (not part of PCS)**	**PCS & PRO (PRO, part of PCS)**
Altmetrics	Higher	Lower	Highest
Citation impact	Lower	Higher	Highest

The results of the analyses are shown in Figure 1, which displays MHq' values for PCS, PRO, and PCS & PRO with upper and lower bounds of 95% confidence intervals (CIs). The results for the traditional metric "citations" are in accordance with expectations. The average citation impact for PRO is significantly higher than that for PCS (PCS & PRO is on a similar level as PRO). For altmetrics scores, we have very different results. All results, however, agree with the expectations in Table 2. The altmetrics differ in the extent of impact differences between PCS and PRO. The altmetrics are sorted by these differences in Figure 1.

Figure 1. MHq' values for PCS, PRO, and PCS & PRO separated by different indicators of impact (citations and altmetrics). The altmetrics are sorted by the impact difference between PCS and PRO.



Consistent results are visible for mentions of papers in policy-related documents and Wikipedia. The impact of the papers referenced in case studies (PCS) is (significantly) higher than the impact of papers submitted as output, and this is especially so in regard to the numbers of mentions in policy-related documents. We also see a similar result for papers mentioned in news items: although the difference is statistically significant, the difference between PCS and PRO is in fact to a lesser extent than it was with mentions of papers in policy-related documents and Wikipedia.
Smaller impact differences between PCS and PRO are visible for blogs and Facebook. For Twitter, this difference is very small and the result might speak against the use of such data as an informative indicator for broad, socio-economic impact assessments. Since tweets do correlate with citations on a relatively low level (as the meta-analysis of Bornmann, 2015a, shows), and thus do not appear to reflect academic impact, it remains a question as to what kind of impact analysts believe is being indexed by tweets although this is at present one of the most popular altmetrics.

Discussion

In a recent study, Ravenscroft et al. (2017) focused on the references cited in case studies and correlated the altmetric scores for the references with the REF scores concerning societal impact (but note that the REF scores are at a UOA/UKPRN – UKPRN refers to UK universities – aggregate level and cannot be discerned for individual case studies, let alone references). They used the Altmetric API to append the Altmetric Attention Scores – a weighted count including a broad range of different altmetrics (e.g., tweets and blog mentions) – to the referenced publications in case studies. Ravenscroft et al. (2017) visualized the relationship between REF scores and Altmetric Attention Score and calculated the Pearson correlation coefficient. The close to zero and negative coefficient ($r = -0.0803$) suggests that these scores seem to measure different things. Thus, the convergent validity of altmetrics with another indicator measuring societal impact does not seem to be supported.
We selected a similar approach to Ravenscroft et al. (2017) by comparing case study metadata with altmetrics. However, whereas Ravenscroft et al. (2017) correlated gross REF scores with altmetrics, we compared specific altmetrics and specific citations for individual PCS and PRO publication records. Furthermore, we did not study the Altmetric Attention Score but instead deconstructed this to examine individual types of altmetrics.
We suggested that there should – on the whole – be high altmetrics scores for PCS (convergent validity) and low scores for PRO (discriminant validity). By contrast, our expectations with citations were the converse of this. We did not expect all papers necessarily to conform to this stereotype. Our results reveal that citations and news as well as mentions on Facebook, in blogs, in Wikipedia and in policy-related documents do appear to have a significant convergent and discriminant validity. The results for Twitter also agree with the expected pattern but the insubstantial absolute differences mean that Twitter does not appear to be a valid source of data for assessing societal impact.
References in REF case studies are a good data source for testing the convergent and discriminant validity of altmetrics data for societal impact assessment. Our results point out that – if metrics are intended to be used for such a purpose – mentions in Wikipedia and policy-related documents would seem to be the more suitable. Since both metrics contain many zero counts, therefore, they should be used in combination with the MHq' indicator. Our study further shows that Twitter counts do not seem to be suitable for societal impact measurements.

References

Arthur, T. (2016). Categorizing Policy Document Citations in PlumX. Retrieved March 10, 2017, from http://plumanalytics.com/categorizing-policy-document-citations-in-plumx/

Bar-Ilan, J., Shema, H., & Thelwall, M. (2014). Bibliographic References in Web 2.0. In B. Cronin & C. R. Sugimoto (Eds.), *Beyond bibliometrics: harnessing multi-dimensional indicators of performance* (pp. 307-325). Cambridge, MA, USA: MIT Press.

Bik, H. M., & Goldstein, M. C. (2013). An Introduction to Social Media for Scientists. *PLoS Biol, 11*(4), e1001535. doi: 10.1371/journal.pbio.1001535.

Bornmann, L. (2014). Validity of altmetrics data for measuring societal impact: A study using data from Altmetric and F1000Prime. *Journal of Informetrics, 8*(4), 935-950.

Bornmann, L. (2015a). Alternative metrics in scientometrics: A meta-analysis of research into three altmetrics. *Scientometrics, 103*(3), 1123-1144.

Bornmann, L. (2015b). Usefulness of altmetrics for measuring the broader impact of research: A case study using data from PLOS and F1000Prime. *Aslib Journal of Information Management, 67*(3), 305-319. doi: 10.1108/AJIM-09-2014-0115.

Bornmann, L., & Haunschild, R. (in press). Normalization of zero-inflated data: An empirical analysis of a new indicator family and its use with altmetrics data. *Journal of Informetrics*. doi: 10.1016/j.joi.2018.01.010.

Bornmann, L., Haunschild, R., & Marx, W. (2016). Policy documents as sources for measuring societal impact: How often is climate change research mentioned in policy-related documents? *Scientometrics, 109*(3), 1477–1495. doi: 10.1007/s11192-016-2115-y.

Derrick, G. E., Meijer, I., & van Wijk, E. (2014). Unwrapping "impact" for evaluation: A co-word analysis of the UK REF2014 policy documents using VOSviewer. In P. Wouters (Ed.), *Proceedings of the science and technology indicators conference 2014 Leiden "Context Counts: Pathways to Master Big and Little Data"* (pp. 145-154). Leider, the Netherlands: University of Leiden.

Digital Science. (2016). *Publication patterns in research underpinning impact in REF2014: A report to HEFCE by Digital Science*. London, UK: Digital Science.

Fleiss, J., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3. ed.). Hoboken, NJ, USA: Wiley.

Haunschild, R., & Bornmann, L. (2017). How many scientific papers are mentioned in policy-related documents? An empirical investigation using Web of Science and Altmetric data. *Scientometrics, 110*(3), 1209-1216. doi: 10.1007/s11192-016-2237-2.

Haustein, S., Bowman, T. D., Holmberg, K., Tsou, A., Sugimoto, C. R., & Larivière, V. (2014). Tweets as impact indicators: Examining the implications of automated bot accounts on Twitter. Retrieved November 27, 2014, from http://arxiv.org/abs/1410.4139

Haustein, S., Peters, I., Sugimoto, C. R., Thelwall, M., & Larivière, V. (2014). Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature. *Journal of the Association for Information Science and Technology, 65*(4), 656-669. doi: 10.1002/asi.23101.

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

Mas-Bleda, A., & Thelwall, M. (2016). Can alternative indicators overcome language biases in citation counts? A comparison of Spanish and UK research. *Scientometrics, 109*(3), 2007-2030. doi: 10.1007/s11192-016-2118-8.

National Research Council. (2014). *Furthering America's Research Enterprise*. Washington, DC: The National Academies Press.

Picardi, C. A., & Masick, K. D. (2013). *Research Methods: Designing and Conducting Research With a Real-World Focus*. Thousand Oaks, CA, USA: SAGE Publications.

Priem, J. (2014). Altmetrics. In B. Cronin & C. R. Sugimoto (Eds.), *Beyond bibliometrics: harnessing multi-dimensional indicators of performance* (pp. 263-288). Cambridge, MA, USA: MIT Press.

Ravenscroft, J., Liakata, M., Clare, A., & Duma, D. (2017). Measuring scientific impact beyond academia: An assessment of existing impact metrics and proposed improvements. *PLOS ONE, 12*(3), e0173152. doi: 10.1371/journal.pone.0173152.

Ringelhan, S., Wollersheim, J., & Welpe, I. M. (2015). I Like, I Cite? Do Facebook Likes Predict the Impact of Scientific Work? *PLoS ONE, 10*(8), e0134389. doi: 10.1371/journal.pone.0134389.

Rons, N. (2012). Partition-based Field Normalization: An approach to highly specialized publication records. *Journal of Informetrics, 6*(1), 1-10. doi: 10.1016/j.joi.2011.09.008.

Rons, N. (2014). Investigation of Partition Cells as a Structural Basis Suitable for Assessments of Individual Scientists. In P. Wouters (Ed.), *Proceedings of the science and technology indicators conference 2014 Leiden "Context Counts: Pathways to Master Big and Little Data"* (pp. 463-472). Leider, the Netherlands: University of Leiden.

Serrano-López, A. E., Ingwersen, P., & Sanz-Casado, E. (2017). Wind power research in Wikipedia: Does Wikipedia demonstrate direct influence of research publications and can it be used as adequate source in research evaluation? *Scientometrics, 112*(3), 1471-1488. doi: 10.1007/s11192-017-2447-2.

Sugimoto, C. R., Work, S., Larivière, V., & Haustein, S. (2017). Scholarly use of social media and altmetrics: A review of the literature. *Journal of the Association for Information Science and Technology, 68*(9), 2037-2062. doi: 10.1002/asi.23833.

Thelwall, M. (2017). *Web Indicators for Research Evaluation: A Practical Guide*. London, UK: Morgan & Claypool.

Thorngate, W., Dawes, R. M., & Foddy, M. (2009). *Judging merit*. New York, NY, USA: Psychology Press.

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., . . . Johnson, B. (2015). *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. Bristol, UK: Higher Education Funding Council for England (HEFCE).