
Article:
Thelwall, M. orcid.org/0000-0001-6065-205X, Kousha, K. orcid.org/0000-0003-4827-971X, Abdoli, M. orcid.org/0000-0001-9251-5391 et al. (4 more authors) (2023) Do altmetric scores reflect article quality? Evidence from the UK Research Excellence Framework 2021. Journal of the Association for Information Science and Technology, 74 (5). pp. 582-593. ISSN 2330-1635

https://doi.org/10.1002/asi.24751

Received: 11 December 2022 | Revised: 23 February 2023 | Accepted: 28 February 2023

DOI: 10.1002/asi.24751

RESEARCH ARTICLE

JASIST WILEY

Do altmetric scores reflect article quality? Evidence from the UK Research Excellence Framework 2021

Mike Thelwall | **Kayvan Kousha** | **Mahshid Abdoli** | **Emma Stuart** | **Meiko Makita** | **Paul Wilson** | **Jonathan Levitt**

Statistical Cybermetrics and Research Evaluation Group, University of Wolverhampton, Wolverhampton, UK

Correspondence
Mike Thelwall, Statistical Cybermetrics and Research Evaluation Group, University of Wolverhampton, Wolverhampton, UK.
Email: m.thelwall@wlv.ac.uk

Funding information
Research England, Scottish Funding Council, Higher Education Funding Council for Wales; Department for the Economy, Northern Ireland

Abstract

Altmetrics are web-based quantitative impact or attention indicators for academic articles that have been proposed to supplement citation counts. This article reports the first assessment of the extent to which mature altmetrics from Altmetric.com and Mendeley associate with individual article quality scores. It exploits expert norm-referenced peer review scores from the UK Research Excellence Framework 2021 for 67,030+ journal articles in all fields 2014–2017/2018, split into 34 broadly field-based Units of Assessment (UoAs). Altmetrics correlated more strongly with research quality than previously found, although less strongly than raw and field normalized Scopus citation counts. Surprisingly, field normalizing citation counts can reduce their strength as a quality indicator for articles in a single field. For most UoAs, Mendeley reader counts are the best altmetric (e.g., three Spearman correlations with quality scores above 0.5), tweet counts are also a moderate strength indicator in eight UoAs (Spearman correlations with quality scores above 0.3), ahead of news (eight correlations above 0.3, but generally weaker), blogs (five correlations above 0.3), and Facebook (three correlations above 0.3) citations, at least in the United Kingdom. In general, altmetrics are the strongest indicators of research quality in the health and physical sciences and weakest in the arts and humanities.

1 | INTRODUCTION

Altmetrics are quantitative indicators for research outputs that are not based on traditional citations from journal articles but are usually derived from web sources. They are widely found in publisher websites, often sourced from the Altmetric.com or PlumX data providers, although CrossRef also provides relevant data (Ortega, 2018). Altmetrics can also be found in the free scholarly search engine Dimensions. Most academics seem to be aware of some of them (Aung et al., 2019), testifying to their importance within the scholarly communication ecosystem. Evidence about the information contained in altmetrics is needed for them to be interpreted effectively, however (Haustein, Peters, Bar-Ilan, et al., 2014; Sud & Thelwall, 2014). This is complicated by a lack of quality control for most and the potential for many of them to be gamed or infiltrated by irrelevant data so they should not be used for important

evaluations (Roemer & Borchardt, 2015; Wilsdon et al., 2015; Wouters & Costas, 2012). Nevertheless, they may have value for formative evaluation, if used carefully (Bar-Ilan et al., 2018), including for authors seeking early indications of likely future impact for individual articles.

The rationale for citation-based impact indicators is that citations can reflect the cited document influencing the citing document, so citation counts partly reflect scholarly influence or impact. Although perfunctory citations also occur, it is still reasonable to use citation counts as scholarly impact indicators if relatively trivial citations can be ignored as "noise" in the system (Moed, 2006). In contrast, the various altmetrics have been hypothesized to reflect different dimensions of attention or impact, and especially societal impact (Kousha, 2019; Priem et al., 2011). Most also have the advantage of appearing before citation counts, giving earlier evidence of interest or impact. There is substantial evidence that one altmetric, Mendeley reader counts, is a scholarly impact indicator and a partial educational impact indicator for journal articles primarily because Mendeley reader counts correlate moderately or strongly with citation counts for articles in most academic fields (Thelwall & Sud, 2016) and can be used as early scholarly impact indicators (Zahedi et al., 2017). Nevertheless, the value and best interpretations of all other altmetrics are uncertain. Tweeter counts, for example, although having moderate correlations with citation counts in some fields (Costas et al., 2015; Haustein, Larivière, Thelwall, et al., 2014), seem to reflect academic interest and author/publisher dissemination activities in many fields rather than the initially hypothesized public interest (e.g., Lemke et al., 2022), despite most Twitter users being non-academics. Biomedical research might be an exception because this research is widely tweeted by the public (Mohammadi et al., 2018; see also: Haustein, Peters, Sugimoto, et al., 2014).

The reason for the ongoing uncertainty about how to interpret altmetrics is a lack of relevant data. Although there are many ways to partially evaluate altmetrics (Sud & Thelwall, 2014), there is no large-scale systematic evidence of the attention given to, or societal impact of, academic research. Thus, there is no direct way to check which altmetrics can reasonably be claimed to be indicators of these, or in which fields. Given this absence, the most common approach has been to correlate altmetric scores with citation counts, as an indicator of scholarly impact, on the basis that positive correlations would at least indicate that altmetric scores are non-random and scholarly-related to some extent. This is almost a paradox since the value of most altmetrics would be in being different from citation counts, but an overlap could nevertheless be expected for any scholarly-related indicator (Thelwall, 2016). Other methods previously used have included content analyses of individual sources (e.g., tweets: Holmberg & Thelwall, 2014), and predicting future citation counts from early altmetric scores (Akella et al., 2021; Thelwall & Nevill, 2018).

Since citation counts are not direct measures of scholarly impact, a better way to evaluate altmetrics would be to correlate them against peer review quality scores for journal articles. This is more direct and may reveal altmetrics that reflect dimensions of quality not well captured by citations. This is plausible since significance (i.e., impact, whether scholarly, societal or other) is one of the three core components of quality, with the other two being rigor and originality (Langfeldt et al., 2020). Although there have been many departmental level comparisons (e.g., Bornmann et al., 2019; Bornmann & Haunschild, 2018) only one (non peer reviewed) publication has previously compared altmetrics with quality scores at the article level. It correlated a range of altmetrics, including Mendeley reader counts and tweet counts from Altmetric.com, with Research Excellence Framework (REF) expert peer review quality scores for 19,580 journal articles from 2008 in 36 field-based Units of Assessment (UoAs). It found only relatively low correlations with quality scores, with the highest in Clinical Medicine (rho = 0.441) and Biological Sciences (rho = 0.363), (HEFCE, 2015), undermining previous claims for the usefulness of altmetrics. A limitation of the analysis was that Altmetric.com started in 2011, so its data for 2008 may have been incomplete. Moreover, given the relatively low numbers of articles in several UoAs, some results may have been imprecise.

The current article updates the REF2014 technical report with more current REF2021 data on the basis that altmetrics have matured over time and Altmetric.com data may be more comprehensive after 2011. Data maturation is likely because the only year previously analyzed, 2008, precedes Altmetric.com's foundation in 2011 and immediately follows Mendeley's creation in 2007. The primary research question is to assess the overall value of altmetrics. The second research question benchmarks against citation counts, as the most widely used research impact indicator.

- RQ1: How useful are Altmetric.com altmetrics as article-level indicators of research quality in all fields?
- RQ2: How do altmetrics compare to raw and field normalized citation counts as indicators of article research quality in all fields?

2 | METHODS

2.1 | Data

Provisional REF2021 scores from March 2022 for 148,977 journal articles from 2014 to 2020 were supplied by the REF team as part of an unrelated project (Thelwall

et al., 2022). These are either the final scores or with a few article scores changed. This includes many duplicate articles that were scored separately because they were supplied by different authors. For security reasons (to hide the authors' scores and their colleagues' scores), University of Wolverhampton submissions were excluded. Each score had been agreed by two subject experts, usually senior researchers, from one of the 34 UoAs and agreed at the UoA level, with norm referencing within each UoA. Thus, the scores are carefully calibrated expert judgments. Each output was scored as 0 unclassified, 1* recognized nationally, 2* recognized internationally, 3* internationally excellent, or 4* world-leading in terms of originality, significance, and rigor (REF2021, 2021). Since the score 0 could be allocated to a very weak article or a stronger article with a technical noncompliance, the 184 articles with score 0 were removed. Articles without Digital Object Identifiers (DOIs) were also excluded (difficult to match Altmetric.com data), as were articles not in Scopus (needed for the citation RQ). Duplicate articles within a UoA were removed, allocating the remaining article the median score of all copies (randomly rounding up or down when the median was a x.5 fraction). Articles after 2018 were excluded because of insufficient time to attract a stable number of citations (Wang, 2013). Exact numbers for each field and year are in the supplementary materials (https://doi.org/10.6084/m9.figshare.21938234) but there were low numbers for the arts and humanities UoAs. In the unrelated project report (Thelwall et al., 2022), see tab. 3.6.1 for overall duplicate information and fig. 3.2.2 for the approximate distribution of the quality scores in each UoA.

Although interpretations of academic quality vary, the REF defines it in its guidelines for assessors (paragraphs 191 to 193 of: REF2021, 2019) using the three standard dimensions of originality, significance, and rigor (Langfeldt et al., 2020). For example, significance is, "the extent to which the work has influenced, or has the capacity to influence, knowledge and scholarly thought, or the development and understanding of policy and/or practice" (REF2021, 2019). More specific guidelines are given for different areas (Panels). For example, only the broadly physical sciences, maths and engineering guidance mentions, "influence on user engagement" (REF2021, 2019).

The REF2021 scores cannot be shared due to REF data protection policy requiring that they are destroyed as personal data (https://www.ref.ac.uk/faqs/). Whilst the REF team publishes a complete list of outputs (https://results2021.ref.ac.uk/outputs) and department level aggregate scores (https://results2021.ref.ac.uk/), REF policy is that individual output scores are deleted before the aggregate scores are released. People accessing any or all of the scores during the process (even if creating them, including all 1,000+ REF panel members) must agree to keep them confidential and confirm their deletion by email to the designated REF Senior Policy Advisor coordinating the information. Whilst the unavailability of the scores is undesirable for research transparency considerations and data sharing is mandated by some journals, exceptions are usually made for legal or ethical reasons, with the former applying here.

Altmetric scores were obtained for each article 2014–2018 by querying its DOI with the Altmetric API (Robinson-García et al., 2014) in Webometric Analyst (https://lexiurl.wlv.ac.uk/) for Altmetric's public record during April–May 2022. Altmetric was chosen in preference to PlumX for its free API that allowed batch downloading of records for all articles. Articles without a record in Altmetric.com (23.6%) were excluded from the 2014–2018 data. It would also have been reasonable to assume that such articles had altmetric scores of 0, but some may also have had Altmetric records with a different DOI (either configured differently or for a different version of the article, such as a conference paper, preprint, or update). For example, UoA 11 Computer Science and Informatics had a large minority of articles without DOI matches in Altmetric.com. Investigations of these articles found that they sometimes had Altmetric scores associated with an ArXiv DOI for the preprint of the article. Altmetric.com presumably knew the official DOI but used the preprint DOI as the primary source for API queries and the online record. Thus, assuming articles with DOIs without Altmetric.com API query matches would have altmetric scores of 0 would sometimes be false. Nevertheless, removing 23.6% of the articles, with many of these likely to have low scores, is a substantial change.

Since Altmetric's Mendeley data may not be systematically updated for all DOIs, Mendeley records were captured directly from Mendeley using its API, again in Webometric Analyst for each article 2014–2017 (see below for the rationale for excluding 2018) during April–May 2022. For this set, articles without a Altmetric.com record were not excluded. Finally, for comparison, Scopus citation counts from January 2021 for each article 2014–2017 were also added by DOI. Article numbers for the processing stages are in Table 1 and sample sizes are in Table 2.

The Scopus citation counts were converted into Normalized Log-transformed Citation Scores (NLCS) to give a theoretically better citation-based indicator (Thelwall, 2017). Field normalization of citations in scientometrics is common because citation rates vary between fields and the normalization process factors this out (Waltman et al., 2011). NLCS values were obtained by first log-transforming all citation scores with $\log(1 + x)$ to reduce skewing, then

TABLE 1 Sample sizes for the data processing for UoAs 1–34 treated separately.

Set of articles	Journal articles
All REF2021 outputs of all types (e.g., 28,699 books or book parts)	[185,594]
All REF2021 journal articles	152,367
REF2021 journal articles supplied	148,977
With DOI	147,164 (98.8%)
With DOI and matching Scopus 2014–2020 by DOI	133,218 (89.4%)
Not matching Scopus by DOI but matching with Scopus 2014–2020 by title	997 (0.7%)
Not matched in Scopus and excluded from analysis	14,762 (9.9%)
All REF2021 journal articles matched in Scopus 2014–2020	134,215 (90.1%)
All REF2021 journal articles matched in Scopus 2014–2020 except score 0	134,031 (90.0%)
All non-duplicate REF2021 journal articles matched in Scopus 2014–2020 except score 0	122,331 [90.0% effective]
All non-duplicate REF2021 journal articles matched in Scopus 2014–2018 except score 0	87,739 [64.6% effective]
All non-duplicate REF2021 journal articles matched in Scopus 2014–2018 except score 0, with DOI and matching an Altmetric.com record	67,030 (76.4% of above)
All non-duplicate REF2021 journal articles matched in Scopus 2014–2017 except score 0	68,245 [50.2% effective]
All non-duplicate REF2021 journal articles matched in Scopus 2014–2017 except score 0, with DOI [for the Mendeley API]	67,736 (99.3% of above)

Note: Effective percentages ignore duplicate articles.

averaging the log-transformed values separately for each Scopus narrow field and year. Each article NLCS was then calculated as its log-transformed citation count divided by the average for its field and year. An article in multiple fields would instead have its $\log(1 + x)$ divided by the average over all relevant fields. The fields used for this were Scopus narrow fields (Scopus, 2022), which are approximately 325 (depending on year) different categories. An NLCS value of 1 indicates world average citation impact, irrespective of the field and year of the article and the scores are comparable between years and fields.

2.2 | Analysis

Spearman correlations were used to assess the strength of association between REF2021 expert peer review scores and altmetric scores for all research questions. Spearman correlations were used instead of Pearson correlations since citation and altmetric data can be highly skewed (Thelwall & Wilson, 2016). Moreover, the REF scores are ranks. Although they are on a limited scale (four values) and the indicators can have a wide range of numbers, Spearman correlations are appropriate because they test for monotonic relationships, but comparisons between values should be cautious (Thelwall, 2016). Whilst correlation does not show cause-and-effect, positive Spearman correlations suggests that articles with a higher altmetric or citation count tend to have higher REF2021 quality scores. Correlations were calculated separately for each year to reduce the influence of time on the results. For citations counts, at least a 3-year window is usually adequate to get reliable results (Wang, 2013), so only articles from 2014 to 2017 were used for the citation data and, since it is compared to the Mendeley API data, the same was applied to the latter. To keep the 3-year window, articles from 2014 to 2018 were used for the Altmetric.com data collected in 2022. For ease of reporting masses of results, the median correlation across all relevant years was reported, but results for individual years are in the supplementary material (https://doi.org/10.6084/m9.figshare.21938234).

3 | RESULTS

3.1 | Correlations

The correlation results are shown here for the altmetrics supplied by the Altmetric API except those for which the median correlations were below 0.1 for all UoAs: Pinners, Questions, GPlus. The remaining results are displayed in themed batches because there is too much data to fit on one graph. A common scale is used for ease of comparison between graphs.

Both Scopus citations and Mendeley readers have similar levels of correlation with REF2021 provisional quality scores in most UoAs, but the Scopus correlations are higher in all except two (17, 34) (Figure 1). Mendeley readers seem to be particularly weak in the humanities. This might be because Mendeley is a reference manager and humanities reference styles are often based on discussions in footnotes rather than standard format references. Thus, Mendeley is less useful to such scholars and its records may be sparser (Thelwall, 2019). The relatively low correlations for Mendeley in Mathematical Sciences and Computer Science and

TABLE 2 Numbers of journal articles used in the analyses.

Unit of assessment or main panel	2014–2017 articles	2014–2018 articles
1: Clinical medicine	5,735	7,068
2: Public health, health services and primary care	2,259	2,824
3: Allied health professions, dentistry, nursing and pharmacy	5,517	6,216
4: Psychology, psychiatry and neuroscience	4,662	5,499
5: Biological sciences	3,744	4,610
6: Agriculture, food and veterinary sciences	1747	1972
7: Earth systems and environmental sciences	2,205	2,522
8: Chemistry	2063	2,162
9: Physics	3,090	3,333
10: Mathematical sciences	2,922	2037
11: Computer science and informatics	2,504	1749
12: Engineering	10,036	6,236
13: Architecture, built environment and planning	1,280	1,038
14: Geography and environmental studies	1766	2099
15: Archaeology	279	331
16: Economics and econometrics	913	760
17: Business and management studies	5,868	4,847
18: Law	852	877
19: Politics and international studies	1,224	1,417
20: Social work and social policy	1,570	1834
21: Sociology	720	848
22: Anthropology and development studies	469	526
23: Education	1,592	1,666
24: Sport and exercise sciences, leisure and tourism	1,368	1,625
25: Area studies	223	241
26: Modern languages and linguistics	476	366
27: English language and literature	367	261
28: History	535	535
29: Classics	50	28
30: Philosophy	368	348
31: Theology and religious studies	83	59
32: Art and design: History, practice and theory	522	414
33: Music, drama, dance, performing arts, film and screen studies	291	217
34: Communication, cultural and media studies, library and information man	436	465
Main panel A (UoAs 1–6)	21,327	25,239
Main panel B (UoAs 7–12)	22,145	17,353
Main panel C (UoAs 13–24)	17,494	17,392
Main panel D (UoAs 25–34)	3,324	2,905
Total (UoAs 1–34)	67,736	67,030
Total (main panels A to D)	64,290	62,889

Note: Main panel data excludes duplicates within UoAs and the 2014–2018 data excludes articles without Altmetric.com records (24%).

Informatics are presumably due to the LaTeX document formatting language commonly used in these areas (also in parts of Physics), for which Mendeley would be less use. The results confirm that citations and Mendeley readers have the most information value in medicine, health, physical sciences, moderate value in

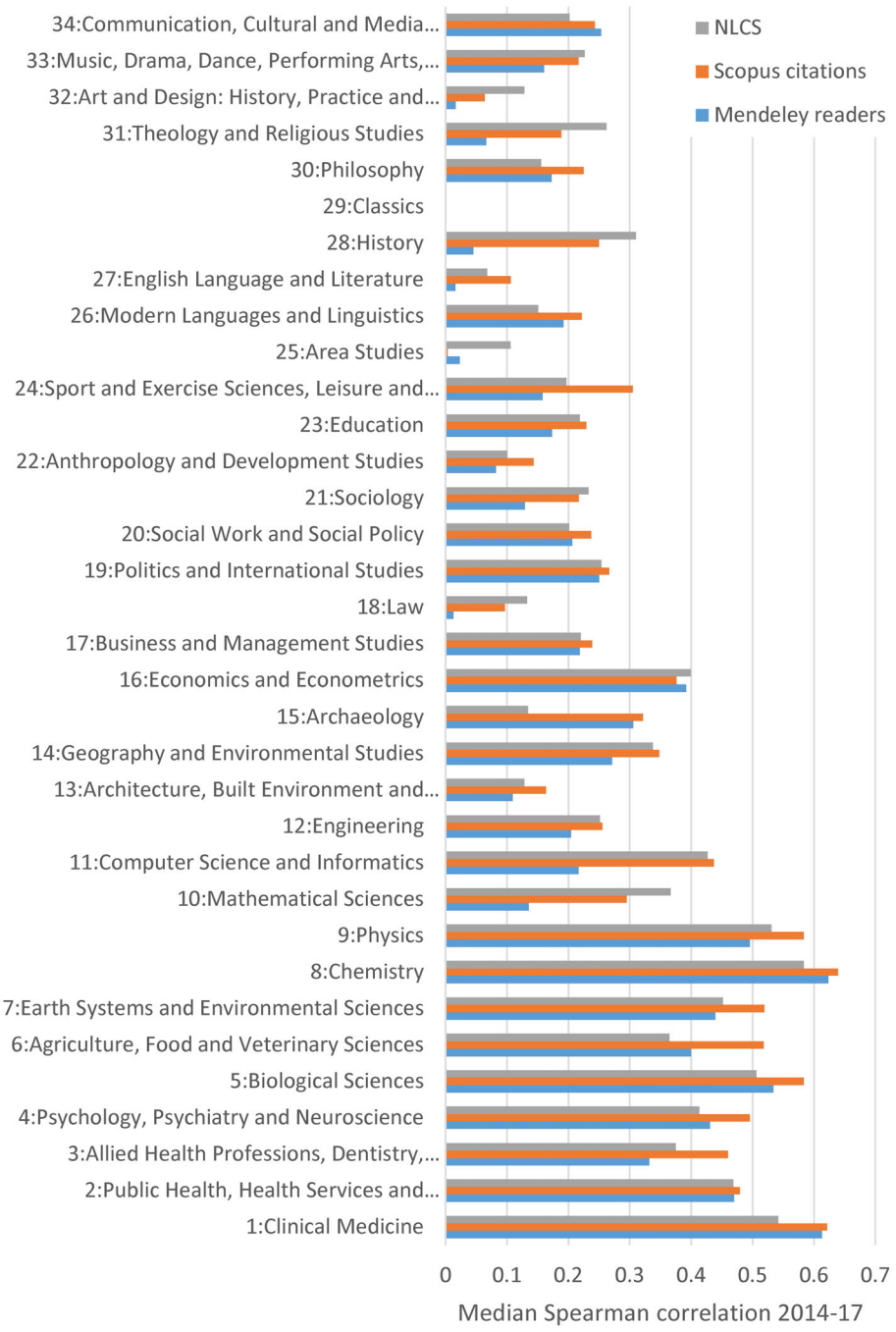

FIGURE 1 Scopus citations (count and NLCS) and Mendeley readers (from Mendeley API) for 2014, 2015, 2016, and 2017 articles: Spearman correlations with provisional REF2021 scores, calculated separately for each UoA and year, with the median across years reported. UoA 29 results have been removed for single figure sample sizes.

mathematics, engineering, and social sciences, and little in the arts and humanities.

Comparing the article quality correlations for the NLCS field normalized citation counts with those for the raw Scopus citation counts, it is surprising that raw citation counts are better indicators of research quality in over two thirds of UoAs (with nine exceptions: 10, 16, 18, 21, 25, 28, 31, 32, 33). This is surprising because field normalized indicators are designed to be fairer than raw citation counts by taking into account the publication field, so an article does not have an advantage for being published in a high citation speciality. In this case the correlations are calculated within field-based UoAs, so field normalization should make little difference. Nevertheless, articles submitted to UoAs by their UK authors can be interdisciplinary or submitted to out-of-field UoA (e.g., because the author is a statistician in a medical department), which the NLCS normalization process should help with. Mostly lower correlations for NLCS suggest that the field normalization process is flawed. This is plausible since Scopus categorizes articles by journal, but article-level classifications more closely align with underlying topics (Klavans & Boyack, 2017). Thus, the results suggest that field normalization, at least based

on articles classified by journal, is usually counterproductive when analyzing articles from a single broad field and year.

The Altmetric.com data for Mendeley gives similar correlations to the Mendeley API data (Figure 2). Altmetric.com claims to have counted readers from CiteULike until December 2014 (Altmetric, 2022b), but its CiteULike data became sparse for 2020 publications (data not used), so it may have ceased collecting new CiteULike data at the end of 2019. Nevertheless, this partial coverage and CiteULike's use by fewer people are the likely causes of lower correlations with REF2021 scores in all UoAs except Area Studies. Combining the CiteULike with the Mendeley counts to give Total Readers does not tend to improve on the Mendeley reader count correlation, so Mendeley readers alone are sufficient.

Of the news related sources, Tweeters (the number of Twitter users tweeting an article URL, although not a complete set: Altmetric, 2022a) seems to be the best indicator of research quality (Figure 3). Nevertheless, Blog and news citations (both from curated lists of sources: Altmetric, 2022a) also have moderate strength as research quality indicators in many UoAs. Facebook Wall links (from a curated list of walls: Altmetric, 2022a) are the weakest, presumably due to smaller numbers of academically-relevant walls curated. Twitter is weaker than Altmetric's Mendeley readers as a research quality indicator in over three quarters of UoAs. The exceptions are mostly in the social sciences, arts, and humanities: UoAs 6, 13, 14, 18, 22, 25, 28, 30, 34.

Reddit mentions, Wikipedia citations and research highlight reviews ("Recommendations of individual research outputs from Faculty Opinions": Altmetric, 2022a) are all weak indicators of research quality in all fields, presumably for their scarcity. Nevertheless, Wikipedia citations have a moderate correlation with research quality in Archaeology and perform well compared to Mendeley Readers and Tweeters in some arts and humanities subjects (Figure 4).

4 | DISCUSSION

The results are limited by the restriction to the United Kingdom and by the articles analyzed being self-selected by academics to be their best work from 2014 to 2020. Thus, the relatively low proportions of weaker research in the sets used for correlation probably reduces the strength of the correlations. In particular, there are few low quality 1* articles and the absence of a substantial proportion of low quality articles that may well score of 0 on all indicators would reduce all correlations. Conversely, since the United Kingdom is a heavy user of social media, including Mendeley, Twitter and Facebook, it is likely that similar correlations would be lower for most other countries. The value of altmetrics may also change over time as the demographics of their users shift. For example, the desktop version of Mendeley started to be phased out in September 2022 (Shlyuger, 2022), which may lose it some users. Another limitation is the use of the REF concept of research quality. Whilst it incorporates the main three quality dimensions, other dimensions or interpretations are also valid, the evaluations are likely to be imperfect because not all articles will have an assessor expert enough to reasonably assess them (Sayer, 2014).

4.1 | Altmetrics

Except for Mendeley and Twitter, the results above are the first reported altmetric correlations with research quality scores and so cannot be compared with prior research. Mendeley correlations are discussed below and Twitter here. Compared to the REF2014 results from 2008 (HEFCE, 2015), the current results are 7–10 years newer and are more robust due to taking a median of several years rather than a single year. Even accounting for this, the Twitter correlations above are surprisingly much stronger than the REF2014 correlations (tab. A54 of: HEFCE, 2015). For 2008, the highest Twitter correlation was 0.23 for Art and Design: History, Practice and Theory, the second highest was 0.17 for Public Health, Health Services and Primary Care, and the remaining correlations were below 0.15, with an average of 0.06. This is only a third of the average correlation above (0.18). Thus, either Altmetric.com's data collection has become more systematic since 2008 or Twitter has changed or matured as a scholarly communication platform (including journal social media policies). The former seems likely because Altmetric.com was founded in 2011 so its Twitter data for 2008 may well have been incomplete (Thelwall et al., 2013).

The relatively high correlations for health-related fields may reflect widespread public interest in potentially impactful medical research (Mohammadi et al., 2018). This increases the amount of altmetric data but also suggests that the public tends to be interested in higher quality research to some extent. This is despite public interest in health research being very topic driven, for example with particular concern for cancer and especially breast cancer (Lewison et al., 2008).

4.2 | Mendeley readers versus Scopus citations

The comparison between Mendeley reader counts and Scopus citation counts above contrasts sharply with the

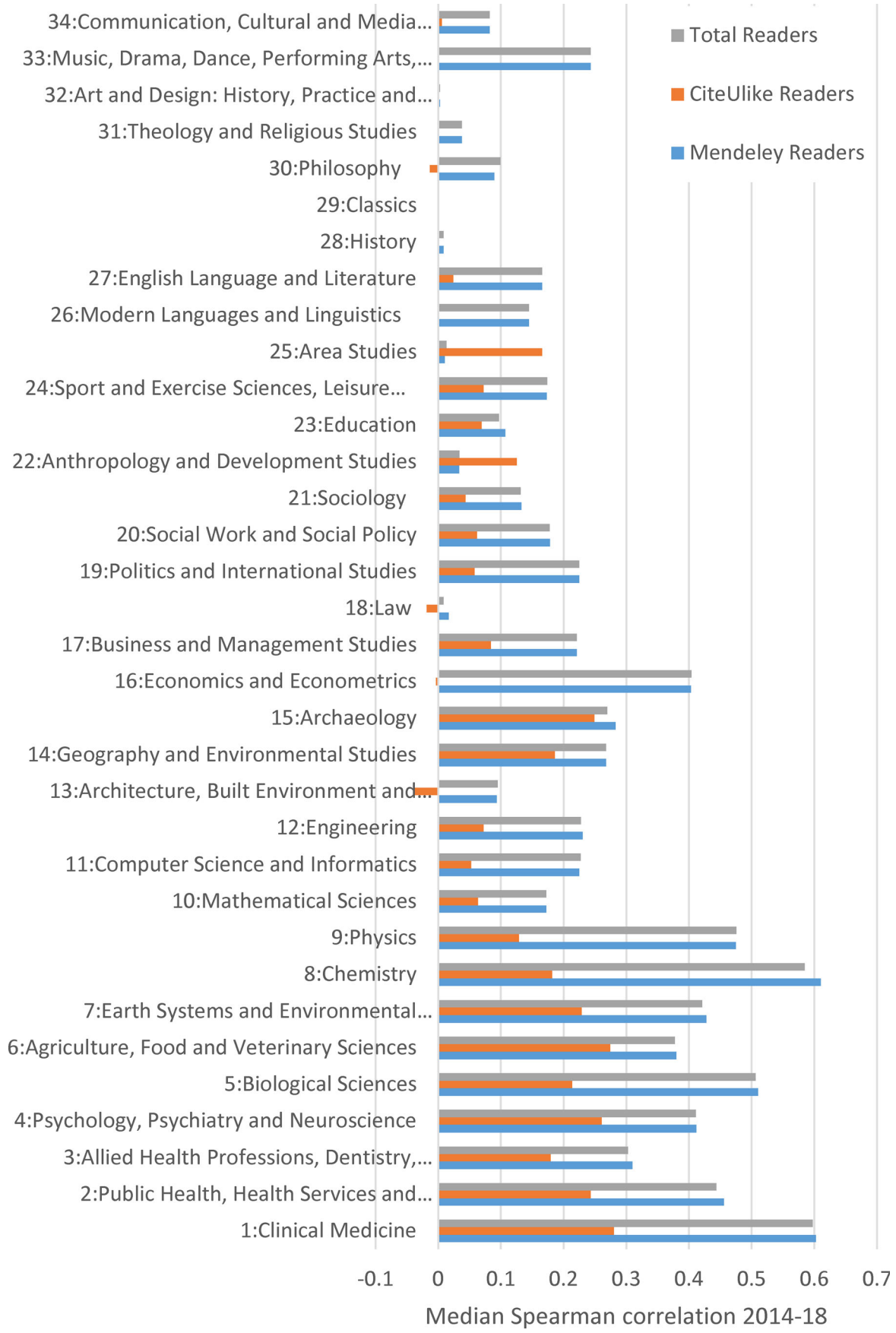

FIGURE 2 Altmetric Scores and online reference manager readers from Altmetric.com 2014, 2015, 2016, 2017, and 2018 articles: Spearman correlations with provisional REF2021 scores, calculated separately for each UoA and year, with the median across years reported. UoA 29 results have been removed for single figure sample sizes.

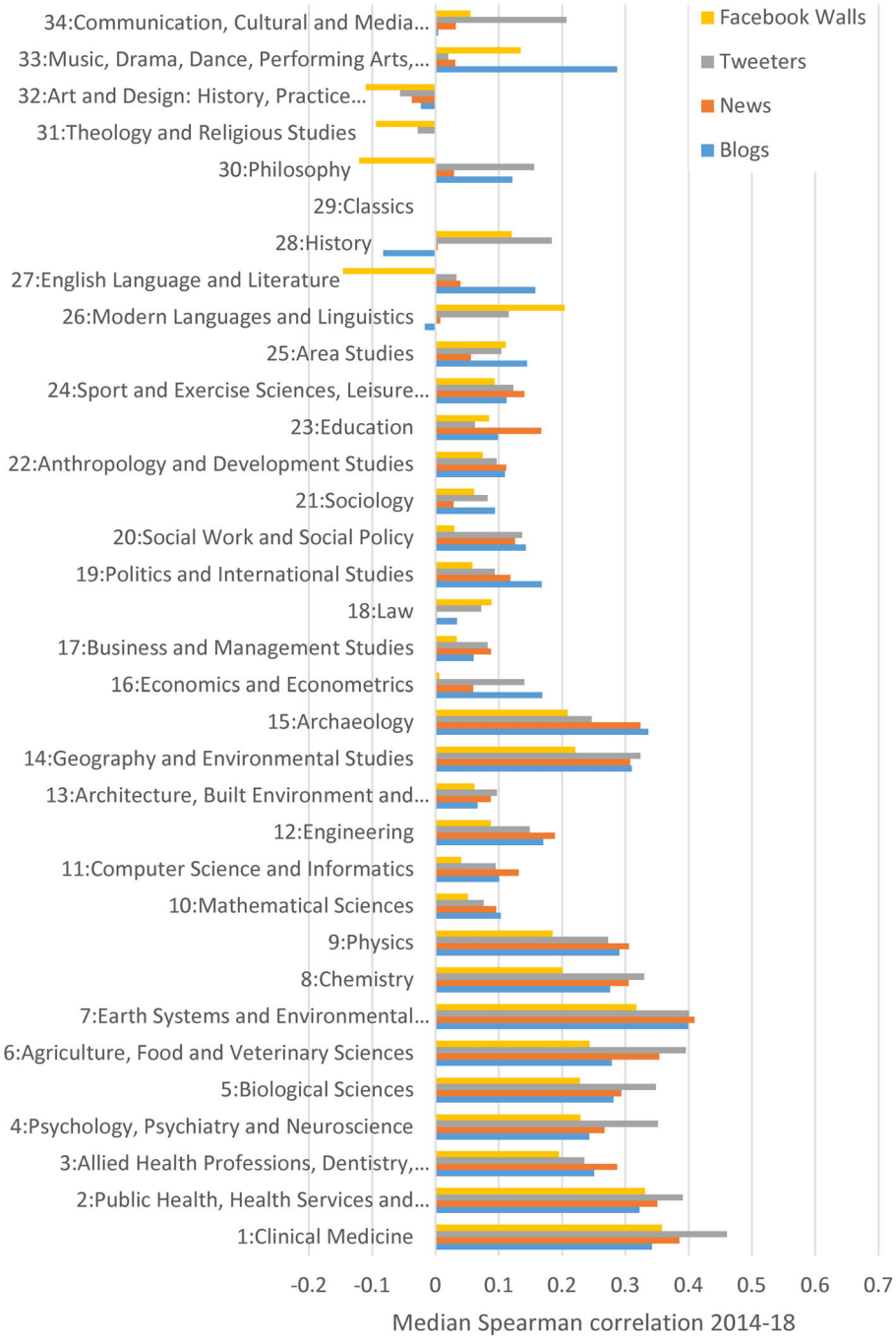

FIGURE 3 Social network and news sites for 2014, 2015, 2016, 2017, and 2018 articles: Spearman correlations with provisional REF2021 scores, calculated separately for each UoA and year, with the median across years reported. UoA 29 results have been removed for single figure sample sizes.

data available from REF2014 (tab. A39 of: HEFCE, 2015). For the 2008 REF2014 data, the Mendeley correlations were overall 52% of the strength of the Scopus citation correlations, with Mendeley being stronger in only 5 out of 36 cases. A likely partial cause of this is Altmetric collecting Mendeley data more systematically now, so its data more closely reflects Mendeley readers. In addition, the first Mendeley results above (Figure 1) use comprehensive data from the Mendeley API. Altmetric previously harvested data from Mendeley for articles that it had registered through other altmetrics, so would have missed some results (Thelwall et al., 2013). Mendeley was launched at the end of 2007 (Henning & Reichelt, 2008) and needed some time to generate a substantial userbase, but its 2008 data nevertheless seems to be as substantial as its later data (Thelwall & Sud, 2016). Thus, probably because of incomplete early Altmetric Mendeley data, the HEFCE analysis seems to have underestimated the value of Mendeley as a research quality indicator. The current results suggest that in most fields outside the arts and humanities it is similar in strength to citation counts in this role, although usually a little weaker.

An exception to the above conclusion is that a previous study claimed that Mendeley readers were as useful

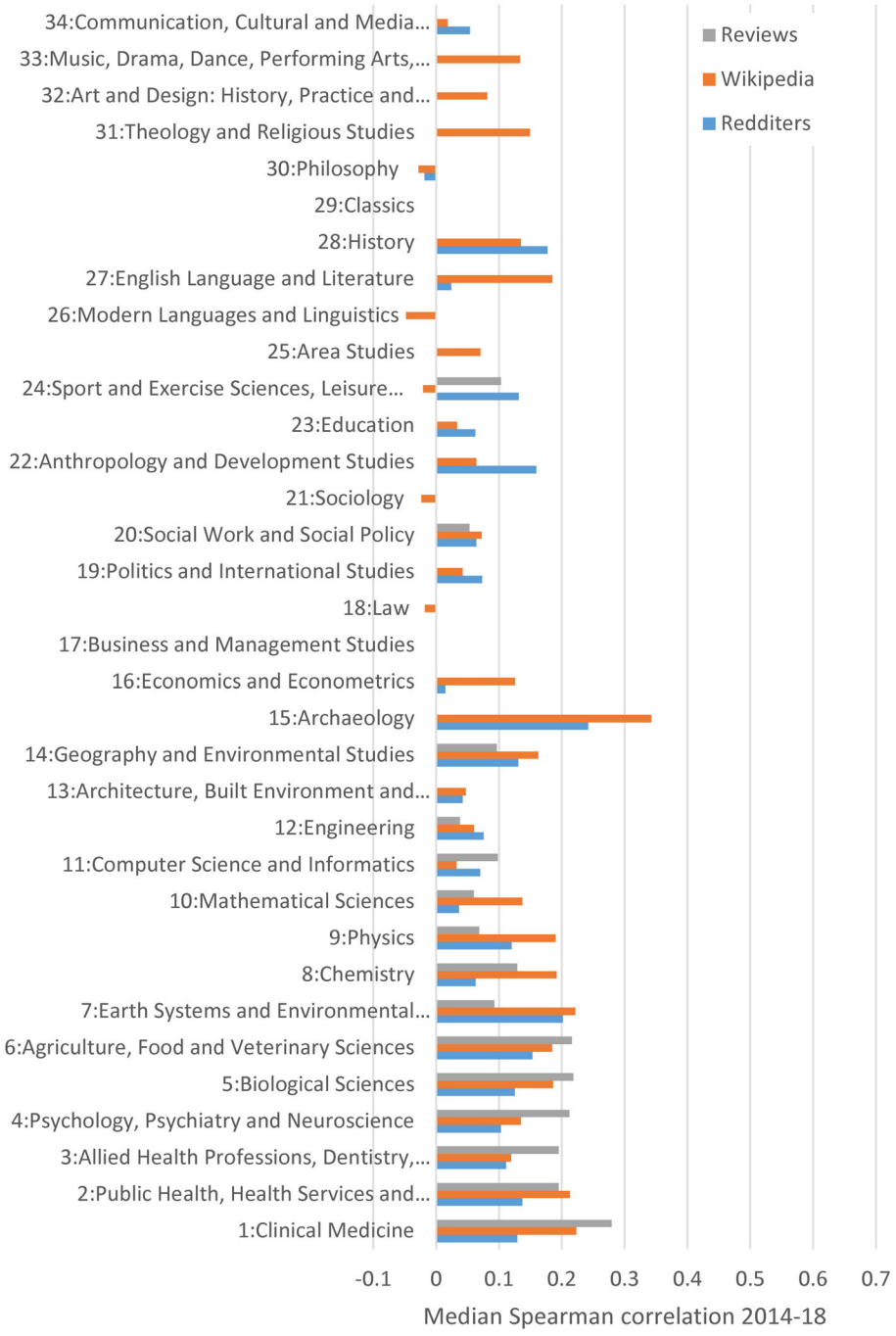

FIGURE 4 Wikipedia, Reddit and facultyopinions.com research highlight reviews for 2014, 2015, 2016, 2017, and 2018 articles: Spearman correlations with provisional REF2021 scores, calculated separately for each UoA and year, with the median across years reported. UoA 29 results have been removed for single figure sample sizes.

in the arts and humanities as elsewhere (Thelwall, 2019). The above results suggest that this is false because Mendeley is of little use in the arts and humanities as a quality indicator. It is even substantially less useful than citations, which are themselves very weak research quality indicators.

4.3 | Field normalization

The comparison between raw Scopus citation counts and field normalized NLCS versions above echo the data available from REF2014, although this was not analyzed in the report (tabs. A3 and A8 of: HEFCE, 2015). For REF2014, Scopus's Field Weighted Citation Impact (FWCI), which is similar to the NLCS above except without the log transformation component (Scopus, 2020), had a stronger correlation with REF2014 final scores for 2008 articles than did raw citation counts in only a third of UoAs (12 out of 36). Thus, similar results for two different field normalized datasets suggest that field normalization of citation counts is not helpful when comparing articles largely within the same broad field (e.g., the 34 REF UoAs), at least when using Scopus narrow fields (approx. 325) for normalization. It is possible that this is due to the log normalization in

NLCS, even though this should improve the robustness of field normalization by avoiding taking the arithmetic mean of highly skewed citation count data for the denominators.

5 | CONCLUSIONS

In answer to the first research question, the Altmetric.com altmetrics that are most useful as article-level indicators of research quality are, in descending order, Mendeley readers, Tweeters, Facebook Walls, News, Blogs, Wikipedia, Reddit and Research Highlights. Of these Mendeley is close to Scopus citations in power as a research quality indicator, and Tweeters is clearly the best of the social web indicators. The last three only have minor value. The evidence is the strongest yet for Mendeley and Twitter and is the first of its kind for the others. The results support the continued use of altmetrics as attention indicators by publishers even though the evidence for some is weak. They particularly strengthen the case for the value of Twitter for article-level altmetrics. Doubt had been previously cast on it due to Twitter's use for publicity and spam, but the current results suggest that these uses have either declined, been filtered out by Altmetric.com, or naturally align with the quality of articles. In terms of field differences, altmetrics have the most value in health fields and the physical sciences, and the least value in the arts and humanities. Nevertheless, none of the correlations are strong enough to claim that altmetrics "measure" research quality in any way. Instead, they are weak or moderate strength *indicators* of research quality, meaning that a high score on them weakly or moderately associates with higher quality, but is far from guaranteeing it, especially given the possibility of manipulation.

For the second research question, none of the altmetrics are as effective as citation counts as research quality indicators, despite Mendeley being a close second. It is reasonable to continue to use Mendeley readers as a substitute for citations as an early impact or quality indicator when the citation window is too narrow for citations, however.

Finally, an accidental by-product of this research was the unexpected finding that field normalizing citations using Scopus narrow fields reduces their value as research quality indicators, at least for the log normalized variant used here, presumably due to problems with the field classifications used. Thus, research evaluators should consider avoiding field normalization when a set of articles to be evaluated are mainly from a single broad field. Alternatively, a more consistent field categorization scheme might be used (Klavans & Boyack, 2017), if available.

ACKNOWLEDGMENTS

This study was funded by Research England, Scottish Funding Council, Higher Education Funding Council for Wales, and Department for the Economy, Northern Ireland. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

ORCID

Mike Thelwall https://orcid.org/0000-0001-6065-205X
Kayvan Kousha https://orcid.org/0000-0003-4827-971X
Mahshid Abdoli https://orcid.org/0000-0001-9251-5391
Emma Stuart https://orcid.org/0000-0003-4807-7659
Meiko Makita https://orcid.org/0000-0002-2284-0161
Paul Wilson https://orcid.org/0000-0002-1265-543X
Jonathan Levitt https://orcid.org/0000-0002-4386-3813

REFERENCES

Akella, A. P., Alhoori, H., Kondamudi, P. R., Freeman, C., & Zhou, H. (2021). Early indicators of scientific impact: Predicting citations with altmetrics. *Journal of Informetrics*, *15*(2), 101128.

Altmetric. (2022a). *Our sources*. https://www.altmetric.com/about-our-data/our-sources-2/

Altmetric. (2022b). *Attention sources coverage dates*. https://help.altmetric.com/support/solutions/articles/6000240455-attention-sources-coverage-dates

Aung, H. H., Zheng, H., Erdt, M., Aw, A. S., Sin, S. C. J., & 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., Haustein, S., Milojević, S., Peters, I., & Wolfram, D. (2018). Peer review, bibliometrics and altmetrics—Do we need them all? *Proceedings of the Association for Information Science and Technology*, *55*, 653–656.

Bornmann, L., & Haunschild, R. (2018). Do altmetrics correlate with the quality of papers? A large-scale empirical study based on F1000Prime data. *PLoS One*, *13*(5), e0197133.

Bornmann, L., Haunschild, R., & Adams, J. (2019). Do altmetrics assess societal impact in a comparable way to case studies? An empirical test of the convergent validity of altmetrics based on data from the UK Research Excellence Framework (REF). *Journal of Informetrics*, *13*(1), 325–340.

Costas, R., Zahedi, Z., & 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.

Haustein, S., Larivière, V., Thelwall, M., Amyot, D., & Peters, I. (2014). Tweets vs. Mendeley readers: How do these two social media metrics differ? *IT—Information Technology*, *56*(5), 207–215.

Haustein, S., Peters, I., Bar-Ilan, J., Priem, J., Shema, H., & Terliesner, J. (2014). Coverage and adoption of altmetrics sources in the bibliometric community. *Scientometrics*, *101*(2), 1145–1163.

Haustein, S., Peters, I., Sugimoto, C., 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.

HEFCE. (2015). *The metric tide: Correlation analysis of REF2014 scores and metrics (Supplementary Report II to the independent review of the role of metrics in research assessment and management)*. Higher Education Funding Council for England https://www.ukri.org/publications/review-of-metrics-in-research-assessment-and-management/

Henning, V., & Reichelt, J. (2008). Mendeley: A last.fm for research? In *2008 IEEE fourth international conference on eScience* (pp. 327–328). IEEE Press.

Holmberg, K., & Thelwall, M. (2014). Disciplinary differences in Twitter scholarly communication. *Scientometrics*, *101*(2), 1027–1042.

Klavans, R., & Boyack, K. W. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology*, *68*(4), 984–998.

Kousha, K. (2019). Web citation indicators for wider impact assessment of articles. In *Springer handbook of science and technology indicators* (pp. 801–818). Springer.

Langfeldt, L., Nedeva, M., Sörlin, S., & Thomas, D. A. (2020). Co-existing notions of research quality: A framework to study context-specific understandings of good research. *Minerva*, *58*(1), 115–137.

Lemke, S., Brede, M., Rotgeri, S., & Peters, I. (2022). Research articles promoted in embargo e-mails receive higher citations and altmetrics. *Scientometrics*, *127*(1), 75–97.

Lewison, G., Tootell, S., Roe, P., & Sullivan, R. (2008). How do the media report cancer research? A study of the UK's BBC website. *British Journal of Cancer*, *99*(4), 569–576.

Moed, H. F. (2006). *Citation analysis in research evaluation*. Springer.

Mohammadi, E., Thelwall, M., Kwasny, M., & Holmes, K. (2018). Academic information on Twitter: A user survey. *PLoS One*, *13*(5), e0197265. https://doi.org/10.1371/journal.pone.0197265

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.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2011). *Altmetrics: A manifesto*. http://www.altmetrics.org/manifesto/

REF2021. (2019). *Panel criteria and working methods (2019/02)*. https://www.ref.ac.uk/publications-and-reports/panel-criteria-and-working-methods-201902/

REF2021. (2021). *Guide to the REF results*. https://ref.ac.uk/guidance-on-results/guidance-on-ref-2021-results/

Robinson-García, N., Torres-Salinas, D., Zahedi, Z., & Costas, R. (2014). New data, new possibilities: Exploring the insides of Altmetric.com. *Profesional de la Información*, *23*(4) http://www.elprofesionaldelainformacion.com/contenidos/2014/julio/03.pdf, 359–366.

Roemer, R. C., & Borchardt, R. (2015). Issues, controversies, and opportunities for altmetrics. *Library Technology Reports*, *51*(5), 20–30.

Sayer, D. (2014). *Five reasons why the REF is not fit for purpose*. https://www.theguardian.com/higher-education-network/2014/dec/15/research-excellence-framework-five-reasons-not-fit-for-purpose

Scopus. (2020). *What is field-weighted citation impact (FWCI)?* https://service.elsevier.com/app/answers/detail/a_id/14894/supporthub/scopus/

Scopus. (2022). *Content—How Scopus works*. https://www.elsevier.com/solutions/scopus/how-scopus-works/content

Shlyuger, D. (2022). *Introducing Mendeley Reference Manager—Designed for today's researcher workflow*. https://blog.mendeley.com/2022/02/22/introducing-mendeley-reference-manager-designed-for-todays-researcher-workflow/

Sud, P., & Thelwall, M. (2014). Evaluating altmetrics. *Scientometrics*, *98*(2), 1131–1143.

Thelwall, M. (2016). Interpreting correlations between citation counts and other indicators. *Scientometrics*, *108*(1), 337–347. https://doi.org/10.1007/s11192-016-1973-7

Thelwall, M. (2017). Three practical field normalised alternative indicator formulae for research evaluation. *Journal of Informetrics*, *11*(1), 128–151. https://doi.org/10.1016/j.joi.2016.12.002

Thelwall, M. (2019). Do Mendeley reader counts indicate the value of arts and humanities research? *Journal of Librarianship and Information Science*, *51*(3), 781–788.

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. (2013). Do altmetrics work? Twitter and ten other candidates. *PLoS One*, *8*(5), e64841. https://doi.org/10.1371/journal.pone.0064841

Thelwall, M., Kousha, K., Abdoli, M., Stuart, E., Makita, M., Wilson, P., & Levitt, J. (2022). *Can REF output quality scores be assigned by AI? Experimental evidence*. arXiv preprint arXiv:2212.08041.

Thelwall, M., & Nevill, T. (2018). Could scientists use Altmetric.com scores to predict longer term citation counts? *Journal of Informetrics*, *12*(1), 237–248.

Thelwall, M., & Sud, P. (2016). Mendeley readership counts: An investigation of temporal and disciplinary differences. *Journal of the Association for Information Science and Technology*, *57*(6), 3036–3050. https://doi.org/10.1002/asi.23559

Thelwall, M., & Wilson, P. (2016). Mendeley readership altmetrics for medical articles: An analysis of 45 fields. *Journal of the Association for Information Science and Technology*, *67*(8), 1962–1972. https://doi.org/10.1002/asi.23501

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. (2011). Towards a new crown indicator: An empirical analysis. *Scientometrics*, *87*(3), 467–481.

Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics*, *94*(3), 851–872.

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). https://core.ac.uk/download/pdf/30612366.pdf

Wouters, P., & Costas, R. (2012). *Users, narcissism and control: Tracking the impact of scholarly publications in the 21st century*. https://www.academia.edu/download/46604436/Users_Narcissism_and_ControlTracking_the20160618-29618-1krd30v.pdf

Zahedi, Z., Costas, R., & Wouters, P. (2017). Mendeley readership as a filtering tool to identify highly cited publications. *Journal of the Association for Information Science and Technology*, *68*(10), 2511–2521.

How to cite this article: Thelwall, M., Kousha, K., Abdoli, M., Stuart, E., Makita, M., Wilson, P., & Levitt, J. (2023). Do altmetric scores reflect article quality? Evidence from the UK Research Excellence Framework 2021. *Journal of the Association for Information Science and Technology*, *74*(5), 582–593. https://doi.org/10.1002/asi.24751