Article:
Kousha, K. and Thelwall, M. orcid.org/0000-0001-6065-205X (2024) Factors associating with or predicting more cited or higher quality journal articles: An Annual Review of Information Science and Technology (ARIST) paper. Annual Review of Information Science and Technology, 75 (3). pp. 215-244. ISSN 2330-1635

https://doi.org/10.1002/asi.24810

Received: 17 February 2023 | Revised: 10 May 2023 | Accepted: 7 June 2023

DOI: 10.1002/asi.24810

REVIEW ARTICLE

ARIST Annual Review of Information Science and Technology | JASIST | WILEY

Factors associating with or predicting more cited or higher quality journal articles: An Annual Review of Information Science and Technology (ARIST) paper

Kayvan Kousha[1] | **Mike Thelwall**[1,2]

[1]Statistical Cybermetrics and Research Evaluation Group, University of Wolverhampton, Wolverhampton, UK

[2]Information School, University of Sheffield, Sheffield, UK

Correspondence
Kayvan Kousha, Statistical Cybermetrics and Research Evaluation Group, University of Wolverhampton, Wolverhampton, UK.
Email: k.kousha@wlv.ac.uk

Funding information
Research England; Scottish Funding Council; Higher Education Funding Council for Wales; Department for the Economy, Northern Ireland as part of the Future Research Assessment Programme

Abstract

Identifying factors that associate with more cited or higher quality research may be useful to improve science or to support research evaluation. This article reviews evidence for the existence of such factors in article text and metadata. It also reviews studies attempting to estimate article quality or predict long-term citation counts using statistical regression or machine learning for journal articles or conference papers. Although the primary focus is on document-level evidence, the related task of estimating the average quality scores of entire departments from bibliometric information is also considered. The review lists a huge range of factors that associate with higher quality or more cited research in some contexts (fields, years, journals) but the strength and direction of association often depends on the set of papers examined, with little systematic pattern and rarely any cause-and-effect evidence. The strongest patterns found include the near universal usefulness of journal citation rates, author numbers, reference properties, and international collaboration in predicting (or associating with) higher citation counts, and the greater usefulness of citation-related information for predicting article quality in the medical, health and physical sciences than in engineering, social sciences, arts, and humanities.

1 | INTRODUCTION

The importance of high-quality research is recognized by those that fund, manage, or conduct it. Nevertheless, there are many small and large decisions that can affect or reflect the quality of academic research and this relationship is not well understood. At one extreme, an international group of governments might discuss whether to fund expensive long-term experimental nuclear reactors or space telescopes, and at the other, a researcher might wonder whether their article title should be phrased as a question or statement. While each decision takes place in a unique context, some occur often enough to be investigated with quantitative methods. Many empirical studies have thus assessed whether common decisions affect, or associate with, the quality or citation impact of academic publications. The purpose of these studies has either been to provide suggestions to help authors or funders to generate the highest quality or impact research or to help evaluators assess the quality or likely future citation impact of published research. Sometimes the direction of causality (i.e., affect vs. reflect research quality/impact) has been unclear. For example, if article titles containing

colons tend to be more cited, is that caused by colon titles being more understandable or informative, therefore attracting extra readers or citers to the article? Or perhaps article titles with colons are more cited because they belong to high citation specialities or journals that expect long, complex article titles? If the former is true then knowledge of the relationship is helpful for authors but in the latter case, it is only helpful for predicting future citations. When cause-and-effect is unclear, authors might use their own judgments about whether to consider a factor.

Predicting future citations for an article or estimating its quality can be useful in a research evaluation context. For example, it may help countries, departments, and research funders to evaluate their research output. Currently, most such evaluations use citation indicators or expert evaluations, with some (e.g., Italy, see below) also exploiting journal-level information. Nevertheless, recent studies have shown that it is possible to go beyond this and make more accurate predictions or estimations with machine learning algorithms that consider other factors, such as authorship team size. This is a potential future application.

Most relevant previous studies have targeted citation counts as proxies for article quality, with the assumption that more cited articles tend to be higher quality. This review discusses separately the minority of studies that target expert review quality scores instead of citation counts. These are particularly important in fields where citation counts are poor proxies for research quality, such as the arts, humanities, and some social sciences.

With a few exceptions, studies of documentary factors associating with the quality or citation impact of research have focused on evidence that can be automatically extracted from the publications, such as title length or the presence of a question mark in a title. They have typically used correlational approaches to find associations, regression to identify associations or make predictions, or machine learning to make predictions. This review summarizes their methods and results, mostly ignoring extra-documentary factors that have also been investigated, such as altmetric attention scores and downloads. The review also ignores discipline-specific factors, such as the influence of hierarchies of evidence in evidence-based medicine, to focus on science-wide issues. The review is split into two connected parts: evidence of associations between document features and citation counts or quality scores; and predicting citation counts or quality scores from document features. The main value of this overview is to identify when factors associating with document quality or impact are universal or show clear patterns, and when these factors are very context dependent and without clear patterns. Unfortunately, the latter is dominant.

2 | METHODS

Scopus alone was used for the literature review searches instead of the Web of Science because it had wider coverage of peer-reviewed journals, includes all core scientometrics journals, and includes most of the Web of Science (Martín-Martín et al., 2021; Singh et al., 2021). Although Google Scholar has wider coverage, it supports less specific queries (e.g., field codes or various Boolean operators) and indexes many low quality journals. The Scopus searches constructed (Table 1) were limited to English language journal articles and reviews, except for queries related to machine learning, for which conference papers are important (Q13 in Table 1). The keywords used in the queries were identified through our prior knowledge of the subject area. Potentially relevant papers were identified through titles and abstracts and then their full texts were downloaded to be reviewed. Many additional studies were also identified from previous meta-analyses or reviews (e.g., Shen et al., 2021; Tahamtan et al., 2016; Xie et al., 2019). Citation chaining in Scopus or Google Scholar, either through checking cited references or citations to relevant studies, was necessary to identify additional papers with keywords not matching the queries such as "Are papers asking questions cited more frequently in Computer Science?" (Fiala et al., 2021), "Easy to read, easy to cite?" (Dowling et al., 2018) or "The quest for citations: Drivers of article impact" (Stremersch et al., 2007).

Some studies had used multiple article text or metadata features to predict citation counts, and a specific Scopus query was used to identify them (see Q9). For Scopus searches with relatively many results (Q6 to Q9), publication years before 2000 were excluded, although earlier key articles were included, when found and relevant (e.g., Glänzel et al., 1995; Katz & Hicks, 1997; Van Raan, 1998).

3 | ARTICLE CONTENT PROPERTIES ASSOCIATING WITH CITATION COUNTS

Many researchers have attempted to model factors that may associate with higher citation counts or have predicted long-term citation counts from journal or article metadata, such as author numbers and country affiliations. This summary focuses on factors intrinsic to a publication rather than external factors, such as peer review scores (e.g., articles with higher reviewer scores tend to be more cited when subsequently published: Bornmann et al., 2012) or altmetrics. This section updates parts of previous ARIST reviews of scholarly communication and bibliometrics (Borgman & Furner, 2002) and scientific collaboration (Sonnenwald, 2007).

This section mostly discusses relatively simple content properties that are under control of the authors, such

TABLE 1 Scopus queries used as part of the process to identify relevant papers.

Factor	Scopus queries to identify relevant papers
Characteristics of article titles and citations	**Q1:** TITLE (title* AND citation* OR cited OR "scientific impact*") AND (LIMIT-TO (SRCTYPE,"j")) AND (LIMIT-TO (DOCTYPE,"ar") OR LIMIT-TO (DOCTYPE,"re")) AND (LIMIT-TO (LANGUAGE,"English")) **Q2:** (TITLE(non-alphanumeric* OR punctuation* OR questionmark* OR "question mark*" OR colon) AND TITLE(citation* OR cited OR "scientific impact*")) AND (LIMIT-TO (SRCTYPE,"j")) AND (LIMIT-TO (DOCTYPE,"ar") OR LIMIT-TO (DOCTYPE,"re")) AND (LIMIT-TO (LANGUAGE,"English"))
Article length and citations	**Q3:** (TITLE (length OR "longer article*" OR "longer paper*" OR "shorter article*" OR "shorter paper*") AND TITLE (citation* OR cited OR "scientific impact*")) AND (LIMIT-TO (SRCTYPE,"j")) AND (LIMIT-TO (DOCTYPE,"ar") OR LIMIT-TO (DOCTYPE,"re")) AND (LIMIT-TO (LANGUAGE,"English"))
Abstract length and citations	**Q4:** (TITLE(abstract*) AND TITLE(citation* OR cited OR "scientific impact*")) AND NOT TITLE("Chemical Abstract*" OR "citations from" OR Abstracting) AND (LIMIT-TO (SRCTYPE,"j")) AND (LIMIT-TO (DOCTYPE,"ar") OR LIMIT-TO (DOCTYPE,"re")) AND (LIMIT-TO (LANGUAGE,"English"))
Article/abstract readability and citations	**Q5:** TITLE (readab* AND citation* OR cited OR "scientific impact*") AND (LIMIT-TO (SRCTYPE,"j")) AND (LIMIT-TO (DOCTYPE,"ar") OR LIMIT-TO (DOCTYPE,"re")) AND (LIMIT-TO (LANGUAGE,"English"))
Cited references and citations	**Q6:** TITLE (reference* AND citation* OR cited OR "highly impact" OR "scientific impact*") AND (LIMIT-TO (SRCTYPE,"j")) AND (LIMIT-TO (DOCTYPE,"ar") OR LIMIT-TO (DOCTYPE,"re")) AND (LIMIT-TO (LANGUAGE,"English")) AND PUBYEAR >1999
Collaboration and citations	**Q7:** (TITLE (collaboration OR "number of author*" OR "co-author*") AND TITLE (citation* OR cited OR "scientific impact*")) AND (LIMIT-TO (SRCTYPE,"j")) AND (LIMIT-TO (DOCTYPE,"ar") OR LIMIT-TO (DOCTYPE,"re")) AND (LIMIT-TO (LANGUAGE,"English")) AND PUBYEAR >1999
Journal impact factor and citations	**Q8:** (TITLE ("impact factor*") AND TITLE (citation* OR cited OR "scientific impact*")) AND (LIMIT-TO (SRCTYPE,"j")) AND (LIMIT-TO (DOCTYPE,"ar") OR LIMIT-TO (DOCTYPE,"re")) AND (LIMIT-TO (LANGUAGE,"English")) AND PUBYEAR >1999
Predicting citation counts of articles	**Q9:** (TITLE (predict* OR factor* OR determin* OR factor* OR characteristic*) AND TITLE (citation* OR cited OR "scientific impact*")) AND NOT TITLE ("impact factor*") AND (LIMIT-TO (SRCTYPE,"j")) AND (LIMIT-TO (DOCTYPE,"ar") OR LIMIT-TO (DOCTYPE,"re")) AND (LIMIT-TO (LANGUAGE,"English")) AND PUBYEAR >1999
Factors associating with research quality and citations	**Q10:** (TITLE-ABS-KEY("Research Assessment Exercise" OR "Research Excellence Framework" OR "Excellence in Research for Australia" OR "Italian research assessment") AND TITLE(citation* OR cited OR "scientific impact*")) AND (LIMIT-TO (SRCTYPE,"j")) AND (LIMIT-TO (DOCTYPE,"ar") OR LIMIT-TO (DOCTYPE,"re")) AND (LIMIT-TO (LANGUAGE,"English")) **Q11:** (TITLE ("research* quality" OR "article* quality" OR "paper* quality" OR "quality of research*" OR "quality of article*" OR "quality of papers*") AND TITLE (citation* OR cited OR "scientific impact*")) AND (LIMIT-TO (SRCTYPE,"j")) AND (LIMIT-TO (DOCTYPE,"ar") OR LIMIT-TO (DOCTYPE,"re")) AND (LIMIT-TO (LANGUAGE,"English")) **Q12:** (TITLE("peer review*" OR "peer-review*") AND TITLE(citation* OR cited OR "scientific impact*")) AND (LIMIT-TO (SRCTYPE,"j")) AND (LIMIT-TO (DOCTYPE,"ar") OR LIMIT-TO (DOCTYPE,"re")) AND (LIMIT-TO (LANGUAGE,"English"))
Machine learning to predict citations	**Q13:** (TITLE("machine learning" OR "deep learning "OR "artificial intelligence" OR AI OR "Gradient Boosting" OR "Random Forest") AND TITLE(citation* OR cited OR "scientific impact*")) AND (LIMIT-TO (DOCTYPE,"ar") OR LIMIT-TO (DOCTYPE,"cp") OR LIMIT-TO (DOCTYPE,"re")) AND (LIMIT-TO (LANGUAGE,"English"))

as title length or the number of references, rather than more complex properties, such as style or research method, which are mentioned at the end. The average numbers of citations per article varies substantially between fields, over time, and between document types, but this issue is not included. Authorship properties are discussed in the next section although many papers have investigated them in parallel with content properties.

Most studies reported here have analyzed document features separately by correlating them with citation counts. For example, a researcher extracted features from article text (title length, number of figures, tables, equations, and characters with no spaces), metadata (number of authors and number of views) and citation counts from high and low cited papers (each 100 articles, $n = 200$) published by MDPI in 2017, finding significant positive associations between citation counts and the number of views, tables, and authors and a negative significant correlation with title length (Elgendi, 2019). A meta-analysis of 262 studies found that there were associations between article "non-scientific features" and citation counts (Pearson's $r < \pm 0.2$). Associations between Journal

Impact Factors (JIFs) or numbers of authors and citation counts had become stronger over time (Mammola et al., 2022). Multiple properties have sometimes been investigated together through a regression model, which has the advantage of assessing their relative contributions. For example, a regression approach would be needed to distinguish between the citation associations of colons and question marks in titles if they tended to be often used together.

No studies have proved cause and effect in the sense of showing that the property investigated influences citation counts. In all cases, it could instead be influenced by a third factor, such as article quality, that also affects citation counts. For example, a more readable abstract might attract more readers and hence more citations. A more readable abstract might also reflect a higher quality article with simple important findings (all the + options in Figure 1), or be a requirement of the top journals in a field. Conversely, in some fields a higher quality article may tend to have a more complex theoretical component with lengthy jargon terms, leading to a less readable abstract. This less readable abstract may attract also more citations by flagging the theoretical terms for academic literature searchers (all the − options in Figure 1). More generally, all combinations of positive negative and neutral relationships in Figure 1 seem possible. There could also be indirect connections between the three factors. For example, better or more cited authors might tend to write more/less readable articles in some fields.

3.1 | Article titles

Interesting, informative, keyword rich, or easy to understand titles may attract the attention of other researchers, making the articles more likely to be found and read and then cited. Many researchers have investigated the relationship between different characteristics of article titles (e.g., length, readability, and the presence of punctuation) and their subsequent citation counts in various subject areas. One unusual study found that articles with more conclusive titles were more likely to be cited for six biomedical topics (Urlings et al., 2021) but most others analyzed simpler properties that could be automatically calculated.

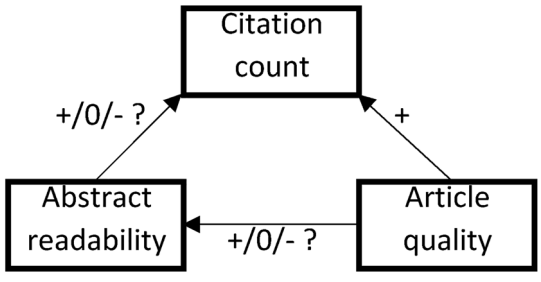

FIGURE 1 Possible relationships between abstract readability, article quality, and citation counts.

3.1.1 | Article title length

Investigations into the relationship between citation counts and article title length, measured in words or characters, have generated mixed results for unknown reasons so there is not a simple and universal relationship between the two. Since journals can have title length restrictions, analyses of multiple journals can find a relationship between title length and citation counts as a side effect of average citation rate differences between journals. Moreover, article types and topics may have different natural title lengths, so any association between citation counts and title lengths can be second order effects of this rather than the title itself influencing citations by attracting readers.

For individual journals, longer titles have been found to associate with more citations for several major general or medical journals in 2005: Lancet, BMJ, Journal of Clinical Pathology, JAMA, Science, and Nature (Habibzadeh & Yadollahie, 2010; Jacques & Sebire, 2010). Insufficient evidence has been found of an association for Addictive Behaviors (Rostami et al., 2014). One of the largest studies investigating the association between article length and citations used 4.3 million papers in articles published 1995–2004 in 1500 large journals, finding that for highly cited journals, shorter titles tend to be more cited (contradicting the above), whereas for the remaining journals, longer titles tend to be more cited (Sienkiewicz & Altmann, 2016).

For datasets containing multiple journals and not analyzing them separately, articles with longer titles tended to be more cited in General Medicine ($n = 6957$) (Van Wesel et al., 2014) and in Biology and Biochemistry ($n = 16{,}058$) and Social Sciences ($n = 15{,}932$) (Didegah & Thelwall, 2013b). Insufficient evidence was found of an association between title length and citations in regression analysis of five major marketing journals (Stremersch et al., 2007) in six PLoS (Public Library of Science) journals (Jamali & Nikzad, 2011), and in Chemistry ($n = 16{,}378$) (Didegah & Thelwall, 2013b). The direction of the association between title length and citation counts has been shown to vary over time for economics articles (Guo et al., 2018). Conversely, shorter titles associated with more citations in articles from 40 psychology journals using structural equation modeling (Subotic & Mukherjee, 2014), in articles from a set of BioMed Central (BMC) and Public Library of Science

(PLoS) journals (Paiva et al., 2012), in articles from both Sociology ($n = 2016$) and Applied Physics ($n = 23,676$) (Van Wesel et al., 2014), and for MDPI journals (Elgendi, 2019). All results in this section could be second order effects of journal title length restrictions, however, for example if higher impact journals encourage shorter titles.

As mentioned above, the most general result for contemporary research is that in highly cited journals, shorter titles tend to be more cited, whereas for less cited journals, longer titles tend to be more cited (Sienkiewicz & Altmann, 2016). Nevertheless, there are exceptions and possibly disciplinary differences and changes over time (Jiang & Hyland, 2022). Also, the precise reason for the association can realistically only be speculated about and may be a second order effect of article type so cause-and-effect is unknown even when a relationship exists.

3.1.2 | Non-alphanumeric title characters

The presence of non-alphanumeric characters in article titles has sometimes been shown to associate with citation counts, presumably because they reflect successful rhetorical styles, such as asking a question or including a subtitle to support both general and specific points. Since the presence of punctuation characters may also associate with title length, any findings on this issue may be second order length effects, unless this is considered. As for title lengths, they may also be second order effects of journal style/impact differences, unless this is also considered.

Articles tend to be more cited when containing a *colon* in each of Lancet, BMJ and Journal of Clinical Pathology (Jacques & Sebire, 2010), but the reverse was true for combined sets of articles from multiple biomedical journals (Jamali & Nikzad, 2011; Paiva et al., 2012), perhaps due to journal mixing. The value of colons is therefore unclear.

Economics articles from multiple journals tend to have 1.6 more citations when they include a *question mark* (Gnewuch & Wohlrabe, 2017). Similarly, computer science journal articles and conference papers (1945–2014) received 16% more citations when including a question mark in their titles (Fiala et al., 2021) and another study supported the association between question marks in the titles and citations (regression coefficient: 0.414) in the field of Software Engineering (Graf-Vlachy et al., 2022). This suggests that questioning titles either attract readers or tend to be associated with citable content, perhaps because the question is answered in the text.

A large-scale analysis of 5% of all Web of Science articles from 1999 to 2008 ($n = 642,807$) found that 68% had at least one out of 29 *non-alphanumeric characters* in their titles, with hyphens, colons, and commas being the most common. In general, articles with non-alphanumeric characters in their titles had higher field-normalized citation impact than titles with only alphanumeric characters. However, there were disciplinary differences, and this association was positive in Clinical Medicine, negative in Biological Sciences, and insignificant in Agriculture and Food Science (Buter & van Raan, 2011).

Overall, the evidence of the relationship between non-alphanumeric characters and citation counts within individual journals is very weak: only available for a few journals and for old studies. For articles from multiple journals, relationships are possible but are likely to vary by field.

3.2 | Article length

Longer papers may tend to attract more citations because they contain more citable content, but some prestigious journals require short articles and so the relationship is not universal. All studies reviewed here measured article length using the total number of pages, even though these depend on page layouts and printing formats and are not relevant to online-only articles. Word counts could be a better indicator of article length but ignore figures, and article full text is needed for such analyses because no major bibliometric database reports word or character counts.

Despite the above caveat, *the evidence is almost unanimous that longer articles tend to be more cited*. This relationship has been found in immunology and surgery (Weale et al., 2004), ecology (Fox et al., 2016), sociology, applied physics, general medicine (Van Wesel et al., 2014), biology and biochemistry, chemistry, mathematics, physics (Vieira & Gomes, 2010), psychology (Haslam et al., 2008), psychiatry (Hafeez et al., 2019), medicine (Falagas et al., 2013), management (Mingers & Xu, 2010), and social sciences (Hodge et al., 2017), as well as for a multidisciplinary set of 1.3 million articles published in 2012 (Haustein et al., 2015). The same positive relationship has also been found for many journals including New England Journal of Medicine, Journal of the American Medical Association, the Lancet (Lyu & Wolfram, 2018), and five economics journals (Hasan & Breunig, 2021). *A meta-analysis* of 18 relevant studies found a moderate positive correlation ($r = 0.310$) between article length and citations (Xie et al., 2019). In contrast, an investigation of Biology and Biochemistry ($n = 16,058$) and Social Sciences ($n = 15,932$) articles from 2000 to 2009 found no significant associations between article length (pages) and citation counts (Didegah & Thelwall, 2013b). Nevertheless, at least two articles have pointed out that expected proportional increase in citations

for longer articles is less than their proportional increase in length (Abt, 1984; Haslam & Koval, 2010).

3.3 | Abstract length

Abstracts have become nearly universal for journal articles over the past half century (Thelwall & Sud, 2022). They help potential readers to understand the topic and results of an article efficiently before they read the full article. Informative abstracts can presumably help relevant research to be quickly identified, and this may be influenced by length, structure, or readability. This section focuses on whether longer abstracts associate with more citations (e.g., because they are more informative) or fewer citations (e.g., because they are harder to digest).

Overall, articles with longer abstracts tend to be more cited. For a million abstracts from eight subject areas, longer abstracts and more sentences in abstracts associated with more citations in all fields (Weinberger et al., 2015). This aligns with similar findings for Biology and Biochemistry, Social Sciences, and Chemistry (Didegah & Thelwall, 2013b). At the journal level this relationship mostly persists: a very large study of 4.3 million papers from over 1500 journals also showed that abstract length positively correlated with citation counts in nearly all journals (Sienkiewicz & Altmann, 2016). In contrast, another large-scale investigation of 300,000 highly cited articles between 1999 and 2008 (30,000 papers per year) found that articles with longer abstracts received fewer citations at the journal level (Letchford et al., 2016), so the overall relationship may be different for highly cited articles.

Positive associations between citation counts and abstract length might be a statistical side-effect of a minority of small articles having very short abstracts. These could be errors (e.g., corrections published as articles), or short articles or comments with a few summary sentences instead of a detailed abstract.

3.4 | Article readability (abstract or full text)

More readable abstracts might be expected to associate with higher citation rates, but the evidence mostly finds the opposite. In some fields, obscurely written abstracts associate with higher impact journals (Tourish, 2020), which might partly explain this finding. A technical problem with assessing the evidence is that there are many ways of measuring readability, such as the relative frequency of rare or long words, with no single best measure. For simplicity, this section treats them all as equivalent.

Articles with more readable abstracts are less cited: For 264,156 articles from five American universities 2000–2009, articles with more readable abstracts were less cited (Gazni, 2011). The same has been found for Biology & Biochemistry (Didegah & Thelwall, 2013b), five major marketing journals 1990–2002 (Stremersch et al., 2007), and 12 emerging technologies (e.g., artificial intelligence [AI], big data, and virtual reality) (Ante, 2022). A large-scale analysis of 4.3 million papers from over 1500 journals also found that articles with more readable abstracts were less cited (Sienkiewicz & Altmann, 2016). The one major exception to these findings is that Economics Letters articles 2003–2012 with more readable abstracts were more cited (Dowling et al., 2018).

In more detail, for 10,000 highly cited and 10,000 uncited English language research articles published during 2008–2017 across 22 subject areas, abstracts of highly cited articles contained more complex, difficult, and professional terms. The highly cited articles also had more adjectives, adverbs, conjunctions, and personal pronouns and longer sentences, making them less readable compared with abstracts of uncited articles (Hu et al., 2021). For 71,628 abstracts from language and linguistics journals (1991 to 2020), abstract readability was low and decreasing over time, and the readability of abstracts negatively correlated with citation counts (Wang et al., 2022). From a different perspective, one investigation introduced an abstract ratio indicator (the sum of repetition of keywords in abstract divided by abstract length), finding that it statistically correlates with citation counts for 5875 articles in Education (Sohrabi & Iraj, 2017, p. 250). Keyword repetition may suggest a narrower focus or an emphasis on a key message. Another study tested five keyword popularity features, finding that keyword popularities can more effectively predict highly cited papers ($n = 746$ articles from 46 journals in marketing and MIS) than can author (author's h-index, publications, or citations) and journal (e.g., JIF and SCImago) features (Hu et al., 2020).

3.5 | Cited references

Citing more references may make articles more visible to researchers using citation tracing in citation databases. Longer articles with more content may also tend to have more references and be cited more. Moreover, higher impact citations may be indications of addressing high citation topics or important issues. There is strong evidence that features of cited references can associate with citation counts, although only for a few fields.

Articles with more references are more cited for Biology and Biochemistry, Chemistry, Mathematics, Physics (Vieira & Gomes, 2010), ecology (Mammola et al., 2021), clinical articles from medical journals (Lokker et al., 2008), AI (Xiao & Jiang, 2020), psychology (Haslam et al., 2008), psychiatry (Hafeez et al., 2019), library and information science (Yu et al., 2014), management (Antonakis et al., 2014), tourism, leisure and hospitality (Cunil et al., 2023), and six biomedical topics (Urlings et al., 2021). This seems to be a universal pattern, perhaps because a longer reference list suggests connections to a wider literature (and therefore potentially more widely relevant), higher quality research (because more justified through references) or a longer article.

Articles with more recent references are more cited for 955,663 Web of Science research articles (Ahlgren et al., 2018) and for 1395 articles across five science and one engineering subjects (Onodera & Yoshikane, 2015). More recent references presumably indicate a more current topic that is more likely to be cited by new articles.

Articles with more high impact references are more cited for 780,049 articles, although there were disciplinary differences (Boyack & Klavans, 2005), for 1.6 million articles (Lancho-Barrantes et al., 2010) and for nanoscience and nanotechnology (Didegah & Thelwall, 2013a), biology & biochemistry, social sciences, and chemistry (Didegah & Thelwall, 2013b). From a related perspective, an investigation of 7749 articles published in 105 journals related to Internet studies found that the authoritativeness of the cited references (the proportion of highly cited references among total cited references in the topic) had a significant positive correlation with citation counts ($\gamma = 0.988$, $p < 0.001$) (Peng & Zhu, 2012). Citing highly cited references may indicate tackling important topics, leveraging ground-breaking prior research, or working within a high citation topic.

Articles with more international references are more cited in nanoscience and nanotechnology (Didegah & Thelwall, 2013a).

3.6 | Other article features

This section reviews article features that have occasionally been investigated for associations with citation counts.

3.6.1 | Images

Analyzing over 4.8 million figures from 650,000 PubMed articles, higher-impact articles had more diagrams per page and a higher proportion of diagrams but a lower proportion of photos (Lee et al., 2017).

3.6.2 | Review articles are more cited

Review articles tend to be more cited than other research articles, although there are some disciplinary differences (e.g., Aksnes, 2006; Colebunders et al., 2014; for a review see Blümel & Schniedermann, 2020). For instance, a very large-scale study of 14.2 million records from Science Citation Index Expanded database during 2000–2015 across 35 science subject areas found that reviews received 1.3–6.7 times more citations than standard research articles, depending on the subject area (Miranda & Garcia-Carpintero, 2018). Citing a review article can be a useful shortcut to reference a body of literature when a detailed analysis is not needed.

3.6.3 | Article findings

For six biomedical research topics within 1990–2018, articles tended to be more cited if they had statistically significant findings, citing research was supportive, there was an empirical research design, the sample size was large, and the funder was commercial (Urlings et al., 2021).

3.6.4 | Article methods

Individual methods may be more cited than average, including questionnaires (Fairclough & Thelwall, 2022), structural equation modeling (Thelwall & Wilson, 2016) and interviews, focus groups and ethnographies, although the degree has changed over time (Thelwall & Nevill, 2021). For biomedical research, methods-focused papers are heavily overrepresented (90%) in the top 100 cited papers (Small, 2018).

3.6.5 | Language

Articles in English or in English-language journals may tend to be more cited (for a review, see Tahamtan et al., 2016), perhaps because English is currently the main international language of scholarly communication, so more scholars can read it. They may also be more cited because a higher proportion of non-English articles address local issues, or because citation indexes mainly index English-language journals (Mongeon & Paul-Hus, 2016), so a greater proportion of other language citations may be lost.

3.6.6 | Open access

Open access (OA) articles seem to have a citation advantage because they are more widely accessible. This is

difficult to check because there are many types of open access, and there are journal-level factors because high- and low-quality journals may be fully OA or fully non-OA. Moreover, it is impossible to account for author decisions, such as if scholars are more likely to ensure that their best work is (or is not) OA. Perhaps because of these factors, together with possible disciplinary differences, current evidence is inconclusive about whether OA advantages exist (Langham-Putrow et al., 2021).

3.6.7 | Topic growth

Articles in a rapidly expanding area, such as a new hot topic, are likely to be more cited than average for the field because the expanding pool of publications has a smaller pool from which to cite (Sjögårde & Didegah, 2022).

4 | AUTHORSHIP TEAM ASSOCIATIONS WITH CITATION COUNTS

This section reviews evidence of associations between authorship team properties and citation counts. Disciplinary differences are particularly likely within this section because of differences in average team sizes and the extent to which equipment and collaboration is essential or beneficial for research.

4.1 | The number of authors

Articles with more authors may tend to be higher quality due to the greater range of expertise or greater challenge of research needing more authors. Nevertheless, larger numbers of authors may also generate more interest for an article through friends and acquaintances, an audience effect (Rousseau, 1992; Wagner et al., 2019), so cause-and-effect is not always clear. While a positive association has been found between citation counts and author counts in nearly all prior studies, there is no agreed formula for the relationship between the two (e.g., linear, logarithmic). Larger authorship teams are likely to involve more institutions and countries, which may alter the relationship between citation counts and team size. Studies of authorship that have taken this into consideration have tended to find a citation advantage of larger teams even after taking these into account.

Many studies of various types have found that articles *with more authors tend to be more cited*, so this seems to be an almost universal (and reasonably strong) phenomenon. This relationship has been found for journal-based studies of eight high impact multidisciplinary (Nature, Science and PNAS), biomedical and science journals 1995–2004 (Hsu & Huang, 2011), for Cell, Science, Nature, New England Journal of Medicine, The Lancet, and JAMA (Figg et al., 2006). Similar results have been found for fields including chemical engineering (Peters & van Raan, 1994), medical sciences (Lokker et al., 2008), psychology (Haslam et al., 2008), pharmacology and pharmacy (Bordons et al., 2013), ecology (Leimu & Koricheva, 2005), library and information science (Sin, 2011), computer science (Ibanez et al., 2013), management (Ronda-Pupo, 2017), biomedical research, chemistry, mathematics (Glänzel, 2002), science and engineering (1955–2000), social sciences (1956–2000), arts and humanities (1975–2000) (Wuchty et al., 2007) the natural and medical sciences, and social sciences and humanities (Larivière et al., 2015), biology and biochemistry, chemistry, mathematics and physics (Vieira & Gomes, 2010), and robotics and AI (Kumari et al., 2020). Similar significant findings have been produced from studies of single countries or institutions, including Norway (Aksnes, 2003), Belgium, Israel, Iran (Chi & Glänzel, 2017), South Africa (Sooryamoorthy, 2009), Italy (Abramo & D'Angelo, 2015; Franceschet & Costantini, 2010), and Harvard University (Gazni & Didegah, 2011). A large study across 27 broad subjects from the 10 countries with most journal articles during 2008–2012 found that increased collaboration associated with more citations for all countries and most subjects, but China and a few fields, including computer science and business, management and accounting had much lower associations between author numbers and citation counts (Thelwall & Maflahi, 2020). There may be a stronger association between citations and research collaboration for developing countries ($r = 0.180$) than for developed countries ($r = 0.112$), however (Shen et al., 2021).

Despite the above positive findings, a few (mostly older) investigations have not found articles with more authors to be more cited, including for eight economics journals in 1990 (Medoff, 2003), chemical articles in 2000 (Bornmann et al., 2012), 14 finance journals 1987–1991 (Avkiran, 1997), nanoscience and nanotechnology 2007–2009 (Didegah & Thelwall, 2013a), and geography and forestry (Slyder et al., 2011). Thus, in specific fields, co-authorship may not associate with more highly cited research. The same seems to be true for monographs (Thelwall & Sud, 2014).

A large investigation across all 27 Scopus broad subjects from 10 countries with the most journal articles during 2008–2012 found that there was a significant increase in the average citation impact of research from single to two authored articles with a subsequent linear rise with additional authorship, giving an *overall logarithm-like shape* (Thelwall & Maflahi, 2020).

4.2 | International collaboration

Internationally co-authored papers tend to attract more citations than domestic articles in most contexts tested so far. This may be due to wider audiences for the research (more people knowing the authors: Wagner et al., 2019), more varied expertise, or more funding (assuming that international collaboration is often triggered by grants). Most investigations of this phenomenon have factored out team size so that internationalism is counted separately from the number of authors.

Articles with international co-authorship receive more citations than articles with domestic co-authorship in Astronomy 1980–1991 (Van Raan, 1998), Sport Sciences 2000–2001 and 2010–2011 (Wang et al., 2015), three computer science sub-fields (big data, machine learning, and data mining) 2005–2019 (Fan et al., 2022) and Biology and Biochemistry and Chemistry 2000–2009 (Didegah & Thelwall, 2013b), in 28 subjects 1977–1986 (Narin et al. 1991), eight subject areas 1996–2012 (Smith et al., 2014), and two broad research areas (Natural and Medical Sciences and Social Sciences and Humanities) (Larivière et al., 2015). Similar patterns have been found at the country level for the United Kingdom 1981–1991 (Katz & Hicks, 1997), Norway 1981–1996 in Natural Sciences (Aksnes, 2003), Europe 2000 (Nomaler et al., 2013), Finland 1990–2008 (Puuska et al., 2014), 35 OECD countries 2003–2013 (Leydesdorff et al., 2019), and for both young ($n = 26$) and old universities (Khor & Yu, 2016).

Against the trend, no citation association with international collaboration has been found for the social sciences 2000–2009 (Didegah & Thelwall, 2013b) and Harvard University 2000–2009 (Gazni & Didegah, 2011). Moreover, some countries may extract more value from international collaboration than others (Lancho-Barrantes et al., 2013; Satish, 2021) and some countries may not benefit from international collaboration, at least in terms of increased citation impact (Smith et al., 2014). For example, American authors in Nature and Science 2004–2008 did have an apparent citation impact increase from international collaboration (Rousseau & Ding, 2016). For biochemistry articles in 2011 ($n = 13{,}578$), research collaboration with the United States associated with increased scholarly impact for published research, whereas co-authorship with some other countries including India and China associated with reduced impact (Sud & Thelwall, 2016).

4.3 | Institutional collaboration

Articles with more institutional affiliations tend to be cited more, perhaps because they are more likely to be funded, or the researchers are more likely to be higher profile to attract extra-institutional collaborators. Since articles with more authors and/or more national affiliations are likely to have more institutional affiliations, most studies have factored out the first two when analyzing the third. Positive associations between the number of institutions and the number of citations have been found for publications affiliated with Harvard University 2000–2009 (Gazni & Didegah, 2011), pharmacology and pharmacy articles by Spanish authors 1998–2000 (Bordons et al., 2013), AI articles 1997–2017 (Fan et al., 2020), nanoscience and nanotechnology articles 2007–2009 (Didegah & Thelwall, 2013a), natural and medical sciences and social sciences and humanities articles 1900–2011 (Larivière et al., 2015), and articles in Cell, Science, Nature, New England Journal of Medicine, The Lancet, and JAMA 1975, 1985, and 1995 (Figg et al., 2006). Some of these studies also showed that the apparent citation advantage of collaboration varied between institution types. Nevertheless, non-significant results have also been found for articles in biology and biochemistry, chemistry, and social sciences 2000–2009 (Didegah & Thelwall, 2013b), and AI article collaborations for some types of research institution (Fan et al., 2020).

4.4 | Author publication and citation records

It seems reasonable to hypothesize that authors with a good track record of publishing or attracting citations would be more likely to write future highly cited papers. It is hard to fully assess this with career-level analyses, but there is some evidence in favor of the hypothesis.

Although the h-index (the largest h such that an author has published at least h articles with at least h citations) is a problematic hybrid indicator because it conflates publishing productivity, citation impact and age, it has often been compared to individual article citation counts, usually with positive results. First author or maximum author h-indexes associate with article citation counts for library and information science (Yu et al., 2014), computer science papers recommended by the China Computer Federation (Qian et al., 2017), astronomy and astrophysics articles published in four journals in 1985 (Wang et al., 2011; Wang et al., 2012), papers written by 65 biomedical researchers (He, 2009), articles in environment and ecology 2006–2007 (Vanclay, 2013), publications by senior researchers from 147 chemistry research groups in the Netherlands 1991–1998 (Van Raan, 2006), and articles in 22 subjects 2000–2009 (Didegah, 2014). There are disciplinary differences in the strength of association, however: a unit increase in the h-index associates with a higher increase in citations in mathematics (6.6%) and

economics & business (5.1%) than in immunology and materials science (both 0.8%) (Didegah, 2014).

From a related perspective, a science-wide analysis of the association between the journal impact (as a proxy for article quality) and authorship properties found that authorship teams publishing more research and higher impact research were more likely to publish in higher impact journals. For this, publishing more cited research was more important than publishing more articles. A first author publishing highly cited research is a science-wide advantage in this regard, while a productive first author is sometimes a disadvantage. A possible explanation is that in some fields, junior first authors might be PhD students conducting particularly careful studies (Thelwall, 2023).

4.5 | Author nationality, institution, and gender

The average citation impact of academic research varies substantially between nations (Confraria et al., 2017). Although there are field differences in this, with countries having high citation specialisms (Elsevier, 2017), essentially richer countries tend to publish more cited work, presumably because of greater infrastructure and resources for research (Confraria et al., 2017). Thus, the national affiliations of the authors of a paper associate with its citation count.

The average citation impact of academic research also varies substantially between institutions within a nation, as evidenced by international citation-based league tables of universities (Waltman et al., 2012). This is likely to be due to some institutions having better researchers and/or more resources and prestige than others. Differences are likely to be greater in countries like the United Kingdom that encourages a hierarchy of universities than in countries like Germany where they are intended to be more equal. The relative citation impacts of universities also vary between specialisms. Thus, the institutions of the authors of a paper associate with its citation count.

Many researchers have found author gender (male vs. female) differences in average citation counts for journal articles, with some studies finding that male first authored articles tend to be more cited (Larivière et al., 2013) and others the reverse (see below). For instance, for over 13,000 research articles and reviews published 2015–2019 in 14 high-impact (greater than 5) general medical journals, the median number of citations per year was 5 for female first authors compared with 6.8 for male first authors (Sebo & Clair, 2023). This issue is complicated by averaging citation counts by the arithmetic mean favoring males whereas averaging citation counts after first taking the natural log, which is statistically better due to the highly skewed nature of citation counts, favoring females (Thelwall, 2018). Using the statistically better approach, the most comprehensive study found a small tendency for female first authored articles to be more cited within the seven English-speaking countries examined 1996–2018 (Thelwall, 2020a), but a follow up analysis of disciplinary differences within six English-speaking countries 1996–2014 found some country/field/year combinations reversing the trend, such as a male citation advantage for Canadian medicine for most years (Thelwall, 2020b).

5 | JOURNAL ASSOCIATIONS WITH CITATION COUNTS

Since the JIF is calculated from the citation rates of the articles in a journal, it is logical to expect articles to be more cited when they are in a journal with a higher JIF. This relationship is not certain, however, since individual highly cited articles may be the cause of a high JIF and the impact factor calculation exclusively counts short term citations.

There is strong evidence from many studies of different fields that *articles in higher impact factor journals tend to be more cited*. This is unsurprising and almost a tautology, as explained above. This has been found for emergency medicine (Callaham et al., 2002), five General & Internal Medicine journals in 2006 (Falagas et al., 2013), six biomedical research topics within 1990–2018 (Urlings et al., 2021), geography and forestry articles (Slyder et al., 2011), social and personality psychology articles from 1998 (Haslam & Koval, 2010), environment and ecology articles 2006–2007 (Vanclay, 2013), demography articles 1990–1992 (van Dalen & Henkens, 2005), biomedicine (Bornmann & Daniel, 2006), clinical systematic reviews and meta-analyses in 2008 (Royle et al., 2013), Norwegian natural sciences 1981–1996 (Aksnes, 2003), immunology and surgery articles (Weale et al., 2004), internal medicine articles 1991–1994 (Fu & Aliferis, 2010), biology and biochemistry, chemistry, mathematics and physics articles (Vieira & Gomes, 2010), internet studies articles (Peng & Zhu, 2012), nanoscience and nanotechnology articles (Didegah & Thelwall, 2013a), biology and biochemistry, chemistry and social sciences articles 2000–2009 (Didegah & Thelwall, 2013b), pharmacology and pharmacy articles (Bordons et al., 2013), F1000 papers 2000–2004 (Bornmann & Leydesdorff, 2015), 33 plastic surgery journal articles 2016–2017 (Asaad et al., 2020), articles from 31 otolaryngology journals 2018–2019 (Hussain et al., 2022) and 780,049 Web of Science articles in 2002–2003 in 17 out of 24 subject areas (Boyack & Klavans, 2005). Some of these

studies suggested that the JIF was the strongest available bibliometric predictor of article citations.

Despite the extensive findings above, a few studies have found insufficient statistical evidence that articles in journals with high impact factors tend to be more cited. These have covered urology (Willis et al., 2011, $n = 200$), ecology (Leimu & Koricheva, 2005, $n = 214$), and gastroenterology and hepatology (Roldan-Valadez & Rios, 2015). A lack of a relationship can be due to small sample sizes or impact factors being skewed by a few highly cited articles.

6 | PREDICTING CITATION COUNTS

This summary focuses on predicting citation counts from document-related factors rather than external factors, such as peer review scores or altmetrics (e.g., early altmetrics predict longer term citation counts: Thelwall & Nevill, 2018).

6.1 | Variations in the inputs, algorithms, and outputs

Regression and machine learning have been used to predict the long-term citation counts of conference papers or journal articles from a wide range of bibliometric and metadata features, with some also extracting extra inputs from article texts using niche corpora. This section includes studies that have reported an accuracy measure, but not studies that have used a prediction method exclusively to assess the strength of predictive factors without reporting on the overall accuracy. The most important dimensions of variation between investigations include the following, which should be considered when evaluating findings.

6.1.1 | Input dataset type

Journal section (Ibáñez et al., 2009), journal (Ibáñez et al., 2009), set of journals (Abrishami & Aliakbary, 2019; Lokker et al., 2008; Robson & Mousquès, 2016; Yu et al., 2014), set of conferences (Cummings & Nassar, 2020; Lee, 2020; Li et al., 2019), field (Ruan et al., 2020; Xu et al., 2019; Zhao & Feng, 2022), random sample from all fields (Akella et al., 2021). More homogeneous sets of documents in dimensions unrelated to citation counts are easier to predict for. In contrast, less homogeneous sets of documents in dimensions related to citation counts are easier to predict for. To illustrate this, a dataset of two different fields with the same citation rate would be harder to predict for than a single field dataset because the properties of the two fields would mix, confusing the algorithm, without giving extra information. Nevertheless, a dataset of two fields with widely different citation rates would be easier to predict for because field differences could be relatively easily identified and then leveraged to predict citation count differences.

6.1.2 | Fields covered

General medicine (Falagas et al., 2013), internal medicine (Fu & Aliferis, 2010), clinical medicine (Lokker et al., 2008), bioinformatics (Ibáñez et al., 2009), high energy physics theory (Chen & Zhang, 2015; Zhao & Feng, 2022), physics (Zhao & Feng, 2022), environmental science and management (Vanclay, 2013), AI (Cummings & Nassar, 2020; Li et al., 2019; Ma et al., 2021), computer and information science (regression: Lee, 2020), library, information and documentation (Ruan et al., 2020; Yu et al., 2014), Markov chains (Xu et al., 2019), mixed (Abrishami & Aliakbary, 2019), or all (Akella et al., 2021). There are wide differences between fields in the accuracy of citation count predictions because of differences in the extent to which input factors systematically associate with higher citation counts.

6.1.3 | Input data range

A third of a year (Falagas et al., 2013) or a single year (Wang et al., 2020; Yu et al., 2014) to 11 years (Chen & Zhang, 2015). Narrower ranges of years generate more powerful predictions due to increased homogeneity, especially if any of the data is not year normalized. Models with multiple years sometimes include the year as an input parameter or have a fixed or long citation window to compensate.

6.1.4 | Size

Eighty-four papers (Saeed et al., 2008) to 420 papers (Ibáñez et al., 2009) to 175,432 papers (Abrishami & Aliakbary, 2019) and 463,348 articles (Zhao & Feng, 2022). The larger the dataset, the more powerful the predictive power, although the increase in power probably decreases with sample size, perhaps with a logarithmic shape. Smaller datasets may be adequate for simple algorithms with few parameters, such as linear regression, but larger datasets are needed for the more complex machine learning algorithms and especially those with larger feature sets.

6.1.5 | Input features

Early citations (Chen & Zhang, 2015; Ma et al., 2021; Ruan et al., 2020; Wang et al., 2020; Yu et al., 2014), citation graph (Cummings & Nassar, 2020; Zhao & Feng, 2022), number of authors (Lee, 2020; Yu et al., 2014), first/all author productivity (Lee, 2020; Yu et al., 2014), first/all author collaboration rates (Lee, 2020), gender (Haslam et al., 2008), first author country income level (Sin, 2011), other basic first/all author properties (Chen & Zhang, 2015; Fu & Aliferis, 2010; Haslam et al., 2008; Lokker et al., 2008; Wang et al., 2020; Yu et al., 2014), complex first/all author capabilities inferred from a matrix analysis of a large document set (Chen & Zhang, 2015; Lee, 2020), field citation rates (Chen & Zhang, 2015), reference count (Ha, 2022; Wang et al., 2020; Yu et al., 2014), reference impact (Boyack & Klavans, 2005), institution properties (Fu & Aliferis, 2010), abstract readability, abstract terms (Fu & Aliferis, 2010; Ibáñez et al., 2009), title terms (Fu & Aliferis, 2010), keywords (Fu & Aliferis, 2010), topics (Robson & Mousquès, 2016), title/abstract sentence semantic representations (Ma et al., 2021), study design (Falagas et al., 2013), journal section (Ibáñez et al., 2009), document type (BinMakhashen & Al-Jamimi, 2022; Ha, 2022), altmetrics (Akella et al., 2021), journal self-citation rate (Ruan et al., 2020), journal impact (Yu et al., 2014), other journal properties (Fu & Aliferis, 2010; Yu et al., 2014), language (Wang et al., 2020), page count (Robson & Mousquès, 2016), title length (Robson & Mousquès, 2016), abstract length (Lokker et al., 2008; Robson & Mousquès, 2016), article length (Ruan et al., 2020), number of figures and tables (Haslam et al., 2008), publication fortnight (Ibáñez et al., 2009), publication month (Ruan et al., 2020), publication year (Robson & Mousquès, 2016), web bookmarks (Saeed et al., 2008), peer review text semantic representation (Li et al., 2019), online ratings (Lokker et al., 2008), number of studies reported (Haslam et al., 2008). This collection shows the huge variety of inputs that have been tested. Some represent fundamental differences of approach (e.g., including early citation information sharpens the focus to citation count prediction, rather than factors associating with higher citation rates) whereas others represent types of information that can only be extracted from deep data processing, and some are relatively speculative. This variety makes it difficult to identify a core set of features needed.

6.1.6 | Feature selection

Yes (Wang et al., 2020) or No (Saeed et al., 2008). Feature selection refers to a procedure to select the most useful features from an initial set. Feature selection can lead to overfitting (exaggerated accuracy statistics, see below) unless it is conducted on each training set independently or separately on a development set. Regression approaches typically do not use feature selection, unless using stepwise regression or another method to identify the most important inputs.

6.1.7 | Algorithms

One or a range to compare, including linear regression (Saeed et al., 2008; Yu et al., 2014), and logistic regression (Fu & Aliferis, 2010; Ibáñez et al., 2009), as well as classical machine learning algorithms like Support Vector Machines (SVM) (Fu & Aliferis, 2010; Wang et al., 2020), random forest (Robson & Mousquès, 2016), naïve Bayes (Ibáñez et al., 2009), neural networks (Abrishami & Aliakbary, 2019; Wang et al., 2020), and deep learning designs (Ma et al., 2021; Xu et al., 2019). Statistical algorithms that rely on identify linear relationships tend to be less powerful, because less flexible, than most machine learning algorithms. These algorithms vary in typical power, with deep learning being particularly promising but requiring the most input data to work well. Statistical approaches usually risk overfitting by not using a separate test set, so the fitting parameter is reported for the same data used to train the model.

6.1.8 | Algorithmic parameter tuning

Before training, only on the training set (Fu & Aliferis, 2010) or not used (Ibáñez et al., 2009). Algorithmic parameter tuning before training adds the risk of overfitting by exploiting information about the data used to evaluate the accuracy of the algorithm.

6.1.9 | Outputs

Citation counts after 1.5 (Boyack & Klavans, 2005), 2 (Akella et al., 2021), 3 (Chen & Zhang, 2015), 4 (Ibáñez et al., 2009), 5 (Ruan et al., 2020), 6 (Falagas et al., 2013), or 14 (Abrishami & Aliakbary, 2019; Ma et al., 2021) years, citations per year (Robson & Mousquès, 2016; Vanclay, 2013), citation ranks (Saeed et al., 2008), if citation threshold exceeded (binary) (Fu & Aliferis, 2010), if highly cited (BinMakhashen & Al-Jamimi, 2022) few, some or many citations in a year (trinary) (C). Citation rates tend to peak several years after publication, with the peak occurrence varying by field, so the long-term citation counts of most papers are probably approximated by n-year citation counts, where n varies between 5 and 10, depending on the field. If

n is too small, then early citations rather than long-term citations might be predicted.

6.1.10 | Accuracy metrics

Percentage correct (Akella et al., 2021), percentage of variance explained (Robson & Mousquès, 2016), AUC (area under the receiver operating characteristic curve, a standard machine learning metric) (Fu & Aliferis, 2010; Lokker et al., 2008), R^2 (Chen & Zhang, 2015; Lokker et al., 2008), mean squared error (Ruan et al., 2020), MSLE (mean square log-transformed error), (Zhao & Feng, 2022), Normalized Discounted Cumulative Gain (Ma et al., 2021), F1 score (Cummings & Nassar, 2020), and rank correlation (Saeed et al., 2008). Accuracy rates are almost never directly comparable between papers even if they report the same metric because variations in the inputs influence the difficulty of the prediction task. It is not possible to correct for this because the relationship between inputs is unknown and there is no reliable measure of task difficulty that could be used for the correction. Comparisons within the same paper on the same dataset can be fully comparable, however.

6.1.11 | Safeguards against overfitting

None, other than separating training and evaluation sets (Fu & Aliferis, 2010), separate development set, pre-declared parameters. Overfitting is the production of optimistic accuracy estimates because the machine learning algorithm is too tailored to the data analyzed. A classic error in machine learning is to evaluate the accuracy of an algorithm on the same dataset used to train it, often leading to greatly exaggerated accuracy. This is routinely avoided now by using a method like 10-fold cross-validation, which builds the algorithm on part of the data and evaluates it on the remainder, doing this 10 times. Even with 10-fold cross-validation, overfitting can still occur in many ways, such as by trying out many variations of algorithms/feature sets/pre-processing steps and reporting only the best or focusing on the accuracy of the best one. If sufficient data are available, then overfitting can be guarded against by using a development set to select the optimal algorithms and parameters, reporting the accuracy of the selected algorithm/parameters on a non-overlapping evaluation dataset.

6.2 | Examples of prediction studies

Most published citation count prediction experiments have focused on articles from a single field or set of journals. Some are described here in detail to illustrate a variety of approaches.

SVM machine learning models were used to predict the future citations of biomedical research (1991–1994) for articles in six high or low JIF medical journals (JAMA, Lancet, NEJM, BMJ, American Journal of Medicine, and Annals of Internal Medicine) matching eight MeSH headings for types of internal medicine. Overall, 3788 documents, 20,005 article text features (article title terms, abstract terms, MeSH terms, publication type), metadata (number of authors and institutions, number articles for first and last authors in the previous 10 years, quality of first author's institution) and citations (number of citations for first and last authors, JIF) were leveraged. The (binary) task was to predict whether an article would reach a given citation threshold (20, 50, 100, or 500) after 10 years. The results gave an accuracy AUC of 0.86–0.92, which was heuristically judged to be "highly predictive." Follow-up analyses with logistic regression (which was less accurate) assessed the value of the different inputs. First author citations had the greatest association (coefficient = 5.75) with articles reaching a citation threshold of 100, followed by the MeSH topic Smoking: mortality (4.22), the JIF (3.32), and last author citations (3.02) (Fu & Aliferis, 2010).

One unorthodox paper generated unusually complex features from a large set of full text preprints. For example, the input "professional knowledge," was derived from a formula based on collaborations with researchers from other topics. It seems to be an indicator of the extent of interdisciplinary collaboration rather than professional knowledge. This article uses a promising approach, but it is hard to evaluate the usefulness of its inputs because they would be impractical for most datasets (lacking full text, or with substantial gaps). The most powerful input seemed to be the average monthly citation count of the papers, with all non-citation inputs having little predictive power (tab. III of Chen & Zhang, 2015).

Other small-scale studies have used machine learning and different article or metadata features to predict citation counts, such as from machine learning conference papers (Cummings & Nassar, 2020; Li et al., 2019), articles from the selected journals (e.g., Wang et al., 2020; Zhao & Feng, 2022) or papers on a specific topic (Xu et al., 2019).

Long-term citation counts can be predicted from early citation counts and/or metadata. An unusual analysis used a large dataset of annual citation counts from 13 years to predict the citation count in the fourteenth year (Abrishami & Aliakbary, 2019). A more standard approach used a neural network to predict the 5-year citation impact of library, information and documentation articles ($n = 49{,}834$) from the Chinese Social

Sciences Citation Index (2000–2013). The study applied multiple features from article text (document type, article length, title length, funding, month of publication, and punctuation in the title), journal (JIF and number of publications in the journal), authors (e.g., number of authors, productivity, previous citations, h-index and number of organizations), references (e.g., number and age of references, self-citations and percentage of different document types in references), and citations (citations in the first or first two years, number of citing journals in the first or first two years), with some positive results (Ruan et al., 2020).

Deep learning is a powerful type of machine learning, although it requires large input datasets and good intuitions about successful network architectures to work well. One investigation used metadata semantic features from AI-related articles published in 20 journals indexed by the China Computer Federation catalog to predict the future citation impact of papers with deep learning techniques for semantic features extraction in the AI subject (Ma et al., 2021). In contrast, another study used multiple altmetric indicators (e.g., Mendeley readers, open peer-review shares, or mentions in Twitter, news or blogs) in addition to other metadata to predict future citations for a random sample of 12,374 articles published in 2015. Using machine learning models, Mendeley readership, maximum followers on Twitter, and academic status (e.g., student, postdoc, researcher, or professor) were the most powerful parameters to predict the short-term and long-term citation impact of papers (Akella et al., 2021).

As the examples above illustrate, citation prediction studies are typically unique and can be radically different in their inputs, methods, and goals. The individual characteristics of each study are limitations that make its results not directly comparable to any other, which greatly complicates the conclusions that can be drawn. This contrasts to the common situation in computational linguistics and information retrieval, for example, where many researchers address the same task on a shared dataset so that their algorithms can be compared (e.g., https://trec.nist.gov/data.html).

6.3 | Summary

It is hard to summarize the situation with machine learning for citation count prediction beyond reporting that this can be done for journal articles and conference papers with many different algorithms and inputs, with a degree of success but that a lack of standardization of all aspects of the task makes it difficult to draw general conclusions about which inputs or algorithms work best, or even what level of accuracy can be expected, however measured. Nevertheless, journal properties, author properties and field/topic properties are all helpful for predicting citation counts, with early citation information being very useful, when relevant to the research goal.

7 | FACTORS ASSOCIATING WITH JOURNAL ARTICLE QUALITY

This section briefly reviews the concept of research quality and reasons why human judgments of it can vary before discussing factors that might associate with article quality. Here, research quality is assumed to be a property that can only be judged by experts, with citation counts (as analyzed above) at best an indictor of it. Thus, with one caveated exception (journal impact as a proxy), all studies of research quality discussed below have used scores derived from expert review, mostly as part of national research evaluation exercises.

7.1 | Concepts of academic research quality

The quality of academic research, when defined, usually encompasses three dimensions: rigor, originality, and (scholarly and societal) significance (Aksnes et al., 2019; Langfeldt et al., 2020). Each dimension is subjective and varies greatly between fields.

There are important variations between fields in the nature of academic rigor. In theory, every article published will be fully rigorous but in practice there are degrees of rigor because almost all research needs assumptions to be practical. The exception is pure mathematics, which does not have to relate to real world concepts, and its key evidence, the proof, is theoretically fully and definitively checkable. Nevertheless, there are still disagreements on the rigor of mathematical proofs and flawed articles are routinely published (Löwe, 2022). In the humanities, rigor applies primarily to argumentation and might entail a reasonably exhaustive consideration of evidence, possibilities, and alternatives, together with convincing assessments of a variety of sources. Qualitative methods rigor might focus instead on ethical dimensions of human subjects research, and the procedures used to tease themes out of data and understand the likely subjective influences of the author(s). From a technological perspective, requirements might be very specific: construction engineering rigor might include the need for bricks to be baked in the appropriate type of oven. In many fields, rigor probably also involves using suitable statistical tests appropriately. While mistakes are easy to identify in these contexts, it is more difficult to

judge between levels of rigor for methods/approaches that are broadly appropriate.

The originality dimension is also clearly subjective. It depends on what the evaluator is already aware of and could be applied to different aspects of research (methods/approaches, research objects, and objectives). Research significance in some specialties might be reasonably assessed with citation counts, but usually encompasses societal impact and evaluators are unlikely to have sufficient knowledge to reliably judge the extent of societal impact, given the myriad potential impacts and the fact that non-academic pathways to impact are rarely documented.

7.2 | Human judgments of academic research quality

Expert judgments about the quality of academic research may differ, including because quality can be judged from different perspectives (Langfeldt et al., 2020). In addition, work that is judged to be high quality within a field because it contributes to the internally agreed field goals may be less highly regarded in national research evaluations because the field goals are not known or are rejected, for example because they are judged to insufficiently consider societal perspectives by being too theoretical or methodologically problematic.

The problem of assessing article quality in a consistent way is complicated by disciplinary differences in the extent to which the quality of an article can be reliably assigned, in the sense of different experts having a high probability of giving the same score. There are several reasons for this. First, there are differences in the extent to which fields are externally-focused, making research significance more difficult to assess. Second, there are differences between fields in the ease with which rigor can be assessed, due to standardization of procedures or the lack of this (Barker & Pistrang, 2005). More generally, not all fields have a relatively uniform centralized agreement on what constitutes high quality research (Trowler, 2014). For example, while this might be expected from fields organized as conceptually integrated bureaucracies (Whitley, 2000) because of relatively centralized control of reputation allocation, it does not occur for fields with varied objects, objectives and/or methods (dis)organized as fragmented adhocracies (Whitley, 2000). In some senses in between these are polycentric oligarchies (Whitley, 2000), where quality is contested between warring paradigms, such as qualitative v. quantitative or empirical vs. theoretical. Other factors being equal, a much higher rate of agreement on quality scores would be expected from fields with the first of the three organizational types.

Given the above factors affecting human judgments of article quality, imperfect human agreement can be expected for all academic fields and substantially different rates of human agreement between fields. These affect the maximum accuracy that it is achievable for AI systems: if the humans disagree on what constitutes quality, then it is more difficult for AI to learn from their decisions. For practical applications, it is also important to take into account human levels of agreement when evaluating the accuracy of machine learning quality estimation systems (Traag & Waltman, 2019). In addition, if there are large disciplinary differences in the variety and standardization of methods, objects, and objectives within a field, then it is technically harder for AI systems to learn markers of quality because they are more diverse: the patterns to discover are fainter. For example, in health-related fields where randomized control trials are reasonably common and recognized as the most robust method, the AI can be expected to learn this. In contrast, most other fields probably do not have a single named high-quality method so it would be more difficult for the AI to distinguish a quality hierarchy of methods, if there is one. For all these reasons, little can be deduced by comparing AI system accuracies between fields. With this caution, accuracy statistics for AI (including statistical approaches with different training and test sets) in different fields is summarized below.

7.3 | Factors associating with article quality

Higher-quality articles tend to be more cited than others from the same field and year in all fields, at least for UK research, with the highest (and strong) correlations being in health, life sciences and physical sciences and the lowest (and weak) being in the arts and humanities (Thelwall, Kousha, Abdoli, Stuart, Makita, Wilson, & Levitt, 2023a). The overall correlations may hide the fact that citations primarily reflect the impact component of research quality, rather than the soundness and originality dimensions (Aksnes et al., 2019). Nevertheless, bibliometric indicators of quality slightly advantage female first authored research, at least in the United Kingdom, and especially in the social sciences, physical sciences, and engineering (Thelwall et al., 2022b). Overall, however, citation counts are indicators of research quality, with substantial disciplinary differences.

The journal citation rate (surprisingly) also associates with article quality in all fields of science, at least for the United Kingdom. A correlation analysis of REF2014 peer review scores and Elsevier's SNIP (Source Normalized Impact per Paper) journal citation impact indicator

(HEFCE, 2015; see confidence intervals in fig. 1 of Thelwall et al., 2022c) for 2008 articles found three out of 27 fields to have negative correlations, but all fields either had statistically significantly positive correlations or had correlation confidence intervals containing positive values. More conclusive evidence was found with a subsequent larger scale study with a finer-grained journal impact calculation. This found weak (0.11) to moderate (0.43) positive correlations between peer review REF2021 quality scores and average journal citation rates (not the JIF, but a similar type of calculation) for all 27 Scopus broad fields and all except one Scopus narrow fields. The correlations were strongest in the medical and physical sciences (and economics) and weakest in the arts and humanities (Thelwall et al., 2022c).

Articles with more authors tend to be higher quality in some but not all fields, at least in the United Kingdom. There are moderately strong Spearman correlations between author numbers and REF2021 quality scores (0.2–0.4) in medicine and the health, life, and physical sciences, but little or no positive association in engineering and the social sciences. In contrast, there was no evidence of association in the arts and humanities, and the decision sciences seemed to benefit from fewer authors (Thelwall, Kousha, Abdoli, Stuart, Makita, Wilson, & Levitt, 2023d). For the United Kingdom, after controlling for the effect of collaboration, having international (rather than national) co-authors associates with higher quality research in 27 out of the 34 Units of Assessment, with collaboration with other advanced economies being particularly advantageous and collaboration with weaker economies tending to be a disadvantage from a quality perspective (Thelwall et al., 2022d).

Finally, UK articles declaring a funding source tend to be higher quality in all fields, irrespective of team size, and seem particularly advantageous for health fields (Thelwall, Kousha, Abdoli, Stuart, Makita, Font-Julián, Wilson, & Levitt, 2023).

8 | ESTIMATING JOURNAL ARTICLE QUALITY SCORES

Although some attempts to predict the long-term citation counts of documents have used these citation counts as a proxy for quality, they are only an indicator of one aspect of quality, scholarly impact. A few studies have attempted to estimate the quality of scholarly documents more directly. The best way to assess the accuracy of AI predictions of quality scores for individual documents seems to be to compare them with expert human judgments, assuming these judgments to be correct. Machine learning has rarely been used to predict the quality scores of individual articles, with two partial exceptions (for different reasons) and two complete exceptions. Fortunately, all four have been multidisciplinary, allowing analyses of disciplinary differences in prediction accuracy.

A science-wide investigation evaluated 32 different machine learning methods on all Scopus-indexed articles published during 2014–2020 across 326 Scopus narrow subjects to predict the quality of published research, using journal impact as a proxy for quality. Specifically, the objective was to identify whether each article had been published in a journal with the top middle or bottom third of citations per paper. The rationale for this was the assumption that, within each field, higher impact journals tend to publish higher quality articles even though the relationship is imperfect. In addition, this does not take into account that some journals may specialize in high or low citation topics, rather than having differing quality thresholds for articles. The data contained 31,273,062 journal articles from Scopus 2014–2020, split into 2310 separate sets for each year and field combination (after discarding some small sets). Citations from 2021 (the Normalized Log-transformed Citation Score: Thelwall, 2017), collaboration (number of authors and number of country affiliations) and article text (words from the title, abstract, and keywords) were used as inputs for the machine learning process. The study tested 30 different regression and machine learning algorithms, finding that the Gradient Boosting Classifier and Random Forest Classifier machine learning methods had the highest levels of accuracy above the baseline (i.e., percentage correct, subtract the percentage correct by guessing that all have the majority class) (an average of 46% and 45% above the baseline, respectively) for predicting the citation-based journal third of articles using the selected features. Accuracy above the baseline was achieved for all fields, although the lowest rates tended to be in the humanities and mathematics (Thelwall, 2022).

A second investigation predicted genuine expert-assigned article-level quality scores from the UK Research Excellence Framework (REF) 2014, on the scale 1*, 2*, 3*, or 4*. It used thresholds rather than machine learning to predict whether an article had been assigned the highest quality score (4*). The data used to make the thresholds included citations, altmetrics, and journal impact indicators (Table 2). The thresholds in each case seemed to be chosen to ensure that approximately the correct number of articles were predicted to be 4*. This approach was applied across all 36 Units of Assessment (UoAs) in the first year of REF2014, which was 2008 (HEFCE, 2015). Although not the purpose of this test, the data can be converted into accuracy statistics and compared to a baseline strategy of predicting that no articles are 4* (i.e., predicting that all articles fall within the

TABLE 2 Accuracy statistics for article-level predictions of whether a REF2014 journal article from 2008 had a 4* score or not across all 36 UoAs (calculated from the two-way summary tables, such as A53, in: HEFCE, 2015). The baseline is predicting that no article is 4*.

Indicator	Accuracy (%)	Baseline (%)	Accuracy above baseline (%)	Articles
Scopus citation counts	76.4	76.6	−0.2	21,060
Google Scholar citations	76.2	76.4	−0.2	21,055
FWCI (field normalized citations)	75.4	76.1	−0.7	19,580
Highly cited percentiles	79.3	91.1	−11.8	19,675
SNIP (a field normalized JIF variant)	74.6	76.2	−1.6	19,130
SCImago journal rank	74.7	76.1	−1.4	19,245
WIPO patent citations	76.1	96.9	−20.8	21,060
Mendeley readers	74.9	86.4	−11.5	21,050
ScienceDirect downloads	67.2	76.0	−8.8	6990
Scopus full text requests	68.3	76.2	−7.9	21,060
Tweets	74.7	94.2	−19.5	21,055

majority class, 1*–3*, sometimes called the ZeroR classifier; the data are no longer available so this is the most accurate baseline). From this comparison, it is not surprising that all strategies had negative accuracy compared to the baseline, although raw citation count thresholds were closest to achieving a positive result (Table 2). The below-baseline accuracies confirm that selecting all articles exceeding a specified threshold science-wide is a very inaccurate way to identify those that are high quality. Because of the disciplinary differences mentioned above, this simple strategy could have achieved a positive accuracy above the baseline for some UoAs.

A third study combined attributes of the first two studies to make genuine machine learning predictions of article quality, as judged by the expert REF2021 assessors. It made separate predictions for journal articles in each of the 34 REF2021 UoAs, separately by year and combining the earliest years (2014–2018). There were for 84,966 articles for 2014–18 in total, varying between UoAs from 56 in Classics (small UoA with few journal outputs) to 12,511 in Engineering. The rarer quality scores 1* and 2* were combined to give a trinary task: predicting 1* or 2* versus 3* versus 4*. There were 10 bibliometric inputs: field and year normalized article citation count, author count, institution count, country count, first author Scopus article count 2014–2020, first author Scopus average citation rate 2014–2020, any author Scopus average citation rate 2014–2020 (maximum), page count, abstract readability, and journal citation rate. There were also 990 textual inputs, chosen using feature selection after the cross-validation splits (to guard against overfitting). The textual inputs were journal names, words, and sets of consecutive two or three words, all taken from titles, keywords, and abstracts. Trained on 50% of the articles on the 2014–18 datasets, the best algorithms achieved accuracy substantially (20%–42%) above the baseline in 11 UoAs: medicine, health and physical sciences, and economics (Table 3). Accuracy was below 20%, and often close to zero or negative in the arts, humanities, social sciences (except economics), and engineering. Correlations between predictions and actual scores were positive in all fields, but with substantial disciplinary differences (Thelwall et al., 2022a).

A follow-up investigation to the above used the same data but split into the 27 Scopus broad subject categories rather than REF UoAs. Accuracy above the baseline exceeded 20% for four broad fields: Multidisciplinary; Biochemistry, Genetics and Molecular Biology; Chemistry, Physics and Astronomy. Accuracy was below the baseline in three broad fields: Arts and Humanities; Dentistry; Pharmacology and Toxicology. Accuracy was below the baseline for nearly all algorithms in two broad fields: Nursing; Energy. The lower accuracy for Scopus broad fields is probably due to them being based on journals, many of which cover multiple disciplines, so the categories are more mixed than UoAs (fig. 9 of Thelwall, Kousha, Abdoli, Stuart, Makita, Wilson, & Cancellieri, 2023). This points to the importance of having accurate categories when predicting article quality. This is probably more important than when predicting article citation rates.

Since the last two studies improve on the first two, the overall conclusions are based on them. In particular, journal article quality scores can be predicted from a careful but not large set of citation, journal and metadata inputs with substantially above baseline accuracy in medicine, health sciences, physical sciences, and economics. Lower accuracy can be achieved in some social sciences and engineering, but there is little chance of making useful predictions in the arts and humanities. Individual fields may be exceptions, however.

Dataset	Articles 2014–2018	Predicted at 50%	Correlation
1. Clinical Medicine	7274	3637	0.562
2. Public Health, Health Services and Primary Care	2855	1427	0.507
3. Allied Health Professions, Dentistry, Nursing and Pharmacy	6962	3481	0.406
4. Psychology, Psychiatry and Neuroscience	5845	2922	0.474
5. Biological Sciences	4728	2364	0.507
6. Agriculture, Food and Veterinary Sciences	2212	1106	0.452
7. Earth Systems and Environmental Sciences	2768	1384	0.491
8. Chemistry	2314	1157	0.505
9. Physics	3617	1808	0.472
10. Mathematical Sciences	3159	1579	0.328
11. Computer Science and Informatics	3292	1646	0.382
12. Engineering	12,511	6255	0.271
13. Architecture, Built Environment and Planning	1697	848	0.125
14. Geography and Environmental Studies	2316	1158	0.277
15. Archeology	371	185	0.283
16. Economics and Econometrics	1083	541	0.511
17. Business and Management Studies	7535	3767	0.353
18. Law	1166	583	0.101
19. Politics and International Studies	1595	797	0.181
20. Social Work and Social Policy	2045	1022	0.259
21. Sociology	949	474	0.180
22. Anthropology and Development Studies	618	309	0.040
23. Education	2081	1040	0.261
24. Sport and Exercise Sciences, Leisure and Tourism	1846	923	0.265
25. Area Studies	303	151	0.142
26. Modern Languages and Linguistics	630	315	0.066
27. English Language and Literature	424	212	0.064
28. History	583	291	0.141
29. Classics	0	0	—
30. Philosophy	426	213	0.070
31. Theology and Religious Studies	107	53	0.074
32. Art and Design: History, Practice and Theory	665	332	0.028
33. Music, Drama, Dance, Performing Arts, Film and Screen Studies	350	175	0.164
34. Communication, Cultural and Media Studies, Library and Information Management	583	291	0.084

TABLE 3 Pearson correlations between AI predictions and actual scores (1* or 2* vs. 3* vs. 4*) for REF2021 data (averaged across 10 iterations). The predictions are for 2014–18 articles, with 50% used for training and the remainder used for the correlation calculation (tab. 4.1.1.1 of Thelwall et al., 2022a).

9 | ESTIMATING DEPARTMENTAL AVERAGE QUALITY SCORES

A few empirical studies have exploited publicly available university or department quality profiles (e.g., institution-UoA departmental quality profiles for previous REFs and Research Assessment Exercise (RAE) 2008 or departmental numerical/star ratings for RAE 1992/1996/2001) to assess the accuracy of different methods to predict these quality profiles from bibliometric data. The purpose has been to assess whether bibliometrics could inform or replace the time-consuming task of manually reviewing the work in the departments assessed. Almost all have been retrospective studies in the sense of making the predictions after seeing the results and so, working with limited sample sizes (usually under 100 departments or under 200 universities) run the risk of overfitting by reporting successful approaches. Nevertheless, the studies collectively show the fields in which bibliometric predictions are the most reliable and the bibliometric data that tends to be most helpful for making predictions.

9.1 | UK RAE/REF scores and bibliometric indicators

Because of the publicly available REF and RAE UoA departmental level results, the United Kingdom has been the target for most investigations into whether departmental-level quality profiles or scores could be predicted with bibliometrics, although only two have had access to article-level quality scores to help with the predictions.

For REF correlations between bibliometrics and quality profiles, it is important to distinguish between those based on total output scores and average output scores. For RAE 1992, 1996, and 2001, departments were given a single score (1, 2, 3, 4, 5, or 5* in 1992; 1, 2, 3a, 3b, 4, 5, or 5* in 1996 and 2001; 68 UoAs), whereas for RAE2008 (67 UoAs) and the Research Excellence Framework (REF) 2014 (36 UoAs) and 2021 (34 UoAs), individual articles were scored and departments were told how many articles had achieved each score (0, 1*, 2*, 3*, 4*) but not the individual article scores (e.g., a department might know that 20 of the 100 articles submitted scored 4* but not which 20 articles). Institutions sometimes averaged their scores to give an informal Grade Point Average (GPA) to allow comparisons over time and between institutions. In each RAE or REF, institutions could submit work to be assessed to any or all UoAs. REFs primarily assess research outputs (e.g., articles, books, and artworks) but include components assessing the research environment and non-academic impact. Bibliometric studies of the REF have tended to focus on journal articles even though these are a minority in the arts and humanities UoAs.

Many bibliometric studies have analyzed whether departmental scores (REF 1992 to REF 2001) or GPAs (RAE 2008 to REF 2021) could be estimated accurately enough with bibliometrics to be able to dispense with the onerous and extensive post-publication peer review needed for each iteration of the exercise. Almost all studies have reported positive correlations between actual and predicted total or average scores. Correlations for the latter will be lower because larger institutions tend to get higher average scores in the United Kingdom, inflating correlations for total scores. In the list below, if a study reports both correlations for both total and average citations, only the latter is mentioned. While early studies usually started from lists of members of departments and then identified citations to their works, later studies used newer more powerful features of citation databases to identify documents for a department or subject area through specific queries, or queried a citation database for the citations to outputs submitted to the REF/RAE from lists of these documents. Some later studies have also used journal information or regressions with a variety of independent variables to predict departmental REF scores.

9.1.1 | Early departmental studies: RAE 1992 to RAE 2001

For RAE 1992, the *average number of citations per member of staff* for first authored publications (a technical limitation) correlated with departmental RAE ratings for Library and Information Science (LIS) for first-authored journal articles (rho = 0.82, $n = 13$) (Oppenheim, 1995, p. 18), Anatomy (rho = 0.49), Genetics (rho = 0.68), and Archeology (rho = 0.74) (Oppenheim, 1997). A much higher correlation was found for average citations per staff member to self-reported publications (not just first authored) from a subset of LIS departments (rho = 0.95, $n = 7$) (Seng & Willett, 1995). In the field of Business and Management Studies, there was a significant correlation ($r = 0.68$) between the sum of a type of *disciplinary journal impact factor* (Discipline Contribution Scoring) for a department's journal articles and the 1992 RAE rating (Thomas & Watkins, 1998).

For RAE 1996, the *average number of citations per member of staff* received in 1998 correlated highly with departmental RAE ratings for Psychology (rho = 0.90) (Smith & Eysenck, 2002).

For RAE 2001, *the total number of citations received by all members of staff* for publications from the assessment

period 1994–2000 correlated with departmental RAE ratings for Archeology (rho = 0.81) (Norris & Oppenheim, 2003), and Music (rho = 0.81) (Oppenheim & Summers, 2008). For Psychology, the *average citations per researcher* received in 1998 correlated with RAE ratings (rho = 0.85) (Smith & Eysenck, 2002). An analysis of departments with at least 20 Web of Science publications and UoAs with at least 20 departments (28 out of 68 UoAs) found positive, statistically significant correlations between *average citations per paper* and departmental RAE scores across all health and medical subjects except nursing, as well as for all natural, formal, and physical sciences except for pure mathematics. Engineering correlations were mostly low and not statistically significant, except for General Engineering, and the social sciences art and humanities correlations were also mostly low and non-significant, except for Business and Management, Economics and Econometrics and Geography (Mahdi et al., 2008). Since this study used the same method for a wide range of UoAs, it gives the earliest systematic evidence of disciplinary differences in the value of citation counts as indicators of RAE quality, with the hierarchy being essentially *physical sciences* > *medical and health sciences* > *formal sciences* > *engineering and social sciences* > *arts and humanities*. The calculations did not consider field or year differences in citation counts, which would have affected the magnitude of correlations but probably not the relative ordering between disciplinary groups.

Also for RAE 2001 but in contrast to the above studies, *regression analyses* allow multiple inputs to be simultaneously compared, identifying the most powerful predictors of scores. A regression on 4400 submissions to the 2001 RAE Political Science panel found that the mean number of citations to the submitted works was the most significant predictor of the RAE scores for the 69 political science departments (standardized coefficient 0.340). Journal articles were the most significant publication type in predicting RAE outcomes compared with authored books or book chapters (Butler & McAllister, 2009). Similar results were found for Chemistry (Butler & McAllister, 2011).

9.1.2 | Mature departmental studies: RAE 2008 and REF2014

For RAE 2008, *average field and year normalized citation counts for articles* have been shown to strongly associate with departmental GPAs in Physics (rho = 0.57), Biology (rho = 0.57) and Chemistry (0.62) (Mryglod et al., 2013), and to weakly associate with GPAs in Mechanical, Aeronautical and Manufacturing Engineering (rho = 0.18), History (0.38), Sociology and Geography and Environmental Studies (both 0.47) (Mryglod et al., 2013). These scores are broadly consistent with the hierarchy found for RAE 2001.

An analysis of RAE 2008 also compared scores with a *journal-level quality indicator* (the sum of the department's Association of Business Schools journal quality scores, divided by the number of staff) and a departmental size indicators with a regression approach, finding the journal data to be highly predictive of departmental scores in Business and Management (regression coefficient beta = 0.773) and Economics and Econometrics (beta = 0.704) (Taylor, 2011). Most fields do not have recognized journal quality rankings, but another study investigated *prestigious publishers and highly cited journals* instead. The reputations of political science journals and book publishers (as measured by a survey of British political scientists) associated with the departmental proportions of top-rated scholarly outputs in the 2008 RAE. For instance, submitted outputs in top 10 journals based on reputational surveys were moderately correlated with the proportions of 4* (rho = 0.49) and 3* (0.33) ratings, whereas this was negative for 2* and 1* rated research (−0.15 and − 0.43 respectively). The proportions of non-top 20 journals in Political Sciences had significant negative correlations with the proportions of 4* (−0.48) and 3* (−0.35) RAE ratings. Similar associations were found between the proportions of articles in the top 20 journals and RAE ratings. The departmental proportion of monographs from top publishers also associated with higher proportions of 4* (0.78) and 3* (0.42) ratings and lower proportions of 2* (−0.37) and 1* (−0.58) RAE ratings (Allen & Heath, 2013). Also for books, a weak, but significant Spearman correlation (rho = 0.387) was found between the 2008 RAE average ranking scores in Communication, Cultural, and Media Studies for 47 institutions and *average Google Books citations* to the 407 books that they had submitted. Since books tend to be much longer than journal articles, even weak evidence from Google Books citation counts might be helpful to support the peer-review process (Kousha et al., 2011).

High significant correlations have been found between departmental RAE 2008 GPAs and *departmental h and g index scores* in Pharmacy (0.77 and 0.70 respectively). The association was weaker in Library and Information Management (0.40 and 0.38) and in Anthropology this association was negative (Norris & Oppenheim, 2010). For REF 2014, stronger Pearson and Spearman correlations were found between departmental h-indexes and different REF score weightings in Biology (ranging from 0.71 to 0.79), Chemistry (0.71 to 0.83), Physics (0.44 to 0.59), and Sociology (0.53 to 0.62) than with institutional normalized citation impact (ranging from 0.37 to 0.67 in different fields) (Mryglod et al., 2015a, 2015b). A blog post also argued that departmental h-indexes could predict RAE 2014 results in

Psychology (Bishop, 2014). The size-dependent nature of this calculation is problematic for some applications, however, since larger departments have an unfair advantage.

Elsevier found a moderate correlation (0.59) between universities' proportions of 4* outputs (world-leading) in REF 2014 and the proportion of their articles that were in the *global top 5% highly cited*. However, there were large disciplinary differences, with the association being much higher in Biological Sciences, Chemistry, Psychology, Psychiatry and Neuroscience, Business and Management Studies, and Computer Science and Informatics ($r = 0.7$ to 0.75) than in other fields, and the association was very weak in Physics and Clinical Medicine (up to $r = 0.3$) (Jump, 2015). This approach has also given high correlations in an academic study (Traag & Waltman, 2019).

For REF 2014, two-thirds of 2014 REF outputs were matched with Web of Science records (133,469 out of 190,962) and different measures were used to assess the agreement between metric-based departmental rankings and REF peer review departmental rankings. There were very high Pearson correlations (r higher than 0.8) between the percentages of 4* rated submissions and *the percentage of top 10% publications* in Economics and Econometrics, Clinical Medicine, Physics, Chemistry, and Public Health. This association was also relatively high (at least 0.7) in Earth Systems and Environmental Sciences, Psychology, Psychiatry and Neuroscience, and Electrical and Electronic Engineering, Metallurgy and Materials. Overall, the associations between citation metrics and REF scores were higher at the departmental level than at the publication level as reported by the HEFCE study (see above HEFCE, 2015), presumably due to averaging effects. Another investigation suggested that top percentile of most cited papers from the UK universities may substitute for REF peer review in Chemistry, Economics and Econometrics, Business and Management Studies, and Physics (Rodríguez-Navarro & Brito, 2020).

Using data from the REF 2014 and citations from Microsoft Academic Graph, there are relatively high correlations between departmental REF GPA and *median citations per submitted publication* (as matched in Microsoft Academic Graph) in 10 subjects, with the correlations being from 0.67 (Physics) to 0.80 (Chemistry; Biological Sciences) (see tab. 3 of Pride & Knoth, 2018).

A study of the association between REF 2014 GPAs and JIFs in Neuroscience, psychiatry, and psychology found that JIF thresholds could be set so that the proportions of publications ranked 4* and 3* would be 95% and 98% accurate (Al-Janabi et al., 2021).

Using machine learning on citation-based indicators (e.g., total citations and average h-index) and Times Higher Education indicators, an experiment assessed if REF 2014 overall university GPAs could be predicted. For this, 79 and 30 UK universities were divided into training and test sets respectively. The number of Web of Science publications, entry tariff and percentage of students were the most significant predictors (Balbuena, 2018), but the sample sizes used were too small for effective machine learning.

9.1.3 | Article-level evidence to predict departmental averages: REF2021

One team was given access to provisional REF2021 scores for journal articles and used them to develop machine learning algorithms to predict their scores from bibliometric and textual information. It combined these article-level predictions to make department-level predictions with the same AI. It assessed whether half of the older articles published 2014–2018 could be predicted by AI, retaining human peer review for the remaining half of the journal articles 2014–2018, all the journal articles 2019–2020, and all non-journal outputs (e.g., books, artworks, websites, and chapters). With this strategy, the scores of individual departments in some UoAs did not change much, with Pearson correlations for the ten most promising UoAs being from 0.66 to 0.91 (Public Health, Health Services and Primary Care; 0.995 if total scores were used) between average output scores with and without partially replacing humans with AI, as described above. Unfortunately, not all departmental level correlations were reported, but the pattern is probably similar to the related Table 3 above, except with higher correlations due to aggregation effects. Despite the high correlations, smaller departments in all UoAs still had a risk of ranking changes, which REF assessors considered too large to accept AI solutions in this way (Thelwall et al., 2022a, 2022b).

The same study also investigated whether assessors could be given REF predictions and prediction probabilities for articles to help when they were undecided about article scores, finding that this might improve overall accuracy by guiding decisions on difficult cases (Thelwall et al., 2022a, 2022b). A practical problem with the credibility of the AI solution was that universities value the league tables formed for each UoA from GPAs and even small changes in scores could lead to a moderate ranking change for smaller institutions. A second practical issue is that JIF-like journal citation information is needed to make the most accurate predictions, but this partly conflicts with UKRI signing the Declaration on Research Assessment (DORA), which is why the technical solution supporting peer review was not recommended for the United Kingdom.

9.1.4 | Summary

The above studies tended to focus on the potential for bibliometrics to replace or supplement peer review in the UK REF or RAE, rather than the limitations, such as funding shifts between institutions if the scores (rather than rankings) change and the potential for perverse incentives when there is a financial incentive to achieve high bibliometric scores (e.g., moving away from less cited important research topics). None of the results contradict the view that "peer review, despite its flaws and limitations, continues to command widespread support across disciplines. Metrics should support, not supplant, expert judgement" (Wilsdon et al., 2015). For this decision support role, the above evidence suggests individual article scores or departmental quality profiles can be predicted to some degree in most fields of research. The predictions are strongest in medicine and the physical sciences, but weakest in the arts and humanities.

9.2 | Peer review and bibliometrics in other countries

9.2.1 | Evidence from Australia

Australia has previously used journal rankings decided by peer review to inform Excellence in Research for Australia (ERA) national evaluations. Although an early investigation found insufficient evidence of an association between citation-based journal metrics and the four tier ERA rankings of Australian social science journals (Haddow & Genoni, 2010), a medium degree of similarity was later found between three journal citation-based indicators and the expert-based ERA rankings. The Source-Normalized Impact per Paper (SNIP) had the highest Spearman correlation (0.54) with ERA rankings ($n = 11{,}137$), followed by raw impact per paper (0.38) and JIFs (0.37) across 27 Scopus subjects, although there were some disciplinary differences. For instance, in Dentistry, journal-based citation metrics had the highest correlations with ERA journal rankings (0.73, 0.78, and 0.72, respectively), followed by Chemical Engineering, and Veterinary Science, whereas very weak associations were found for Social Sciences (0.41, 0.24, and 0.26) (Haddawy et al., 2016).

9.2.2 | Evidence from Italy

Italy uses an output-based periodic research assessment, first known as the VTR and then the VQR. An investigation of institutional aggregate peer review ratings for academic publications submitted to the VTR and their JIFs found significant medium Spearman correlations for Biology (0.48), Chemistry (0.45), and Economics (0.44), suggesting that there is a degree of similarity between peer review outcomes and journal impact in some fields at the level of institutions (Reale et al., 2007). A large multidisciplinary analysis of over 12,000 research articles across 10 subjects also found significant medium-high Spearman correlations between VTR institutional aggregate peer ratings and institutional aggregate article citations across most fields, including Physics (rho = 0.81), Earth Sciences (0.79), Biology (0.69), and Chemistry (0.6) (Franceschet & Costantini, 2011). Another study found some agreement between citation indicators and VQR peer review ratings for 590 Italian articles in Economics, Management, and Statistics (Bertocchi et al., 2015) and there have been arguments that bibliometrics are preferable to peer-review due to cost savings from the time to perform peer review for Italian research assessment (see Abramo et al., 2009; Abramo & D'Angelo, 2011). Nevertheless, recent evidence from the Italian research assessment exercise found that bibliometrics and peer review had weak associations in science, technology, engineering, and mathematics (Baccini et al., 2020).

9.2.3 | Evidence from the Netherlands

The Netherlands does not have a periodic national REF-like procedure but has alternative methods of assessing research quality, sometimes using bibliometric indicators to inform expert judgment. An early investigation of 56 condensed matter physics programs in the Netherlands found that in general there were positive relationships between a range of publication and impact indicators with peer judgments made by expert physics committees, although the strongest Spearman correlations were found between overall jury ratings and the average number of citations per publication (ranging from 0.51 to 0.68) and the field normalized citation averages (0.46 to 0.58) (Rinia et al., 1998). A later investigation of journal articles from 147 university chemistry research groups in the Netherlands (1991–2000) found that both the h-index and the "crown indicator" (field normalized citation count) for research groups significantly and positively correlated with peer judgments of the research quality of published research (Van Raan, 2006).

9.2.4 | Evidence from Norway

A case study of 34 research groups from a Norwegian university found significant, albeit weak, correlations between expert panel ratings and various citation

metrics, including relative subfield citedness ($r = 0.46$), relative citation rate (0.24) and number of citations per person (0.31) (Aksnes & Taxt, 2004). There are also positive associations between different journal citation indicators (SNIP, Scimago Journal Rank and the raw impact per paper) and Norwegian expert-based assessments of journals and series (Ahlgren & Waltman, 2014).

The results from Australia and Italy show that journal information, whether expert rankings or citation-based indicators, can be informative in research evaluations (although explicitly banned in the UK REF), with the latter having statistical validity in some contexts. Because of DORA concerns, however, this approach seems unlikely to be widely adopted. In contrast, evidence from Italy, the Netherlands and Norway tends to confirm that article-level citation-based indicators can validly have a supporting role in research group evaluations.

10 | CONCLUSIONS

The studies reviewed here show that a wide range of factors derived from article text (e.g., length of articles, titles or abstracts, number or impact of cited references and article readability) might be related to the scientific impact of journal articles or conference papers as reflected by citation counts. However, there are disciplinary differences in almost all the results, often without a general pattern, and some findings could be biased by journal style norms that associate with higher or lower impact factors. Some of the associations also varied over time or between journals. Thus, while there are general trends for some properties, there are no universal laws for most, or too little evidence to speculate about such patterns. An additional risk with text mining to predict citation counts is that it is likely to work best by identifying highly cited topics, predicting higher citation counts for all articles on these topics. A successful prediction model for 1 year might be invalid for the next one due to topic changes, so text mining may need rebuilding each year to identify the new hot topics.

In terms of general trends, it seems that more cited research is likely to have more authors, be published in higher cited journals, be longer, and list more and higher impact references. Other potential factors are more variable between disciplines and/or countries, including international collaboration and inter-institutional collaboration. These tend to associate with higher citation counts but there are many exceptions. Moreover, there does not seem to be a general pattern in the association between title and abstract properties and citation counts.

In parallel to the above higher quality research tends to be more cited and in more cited journals, especially in medicine, health, and physical sciences. Other potential factors are more variable between disciplines and/or countries, including author numbers and international collaboration.

The associations found rarely have a clear cause-and-effect relationship. For example, it is not clear whether team size associates with more cited research because larger team research is intrinsically better, funders often insist on large teams, or better researchers find it easier to attract collaborators. Thus, even the clearest findings are only suggestions about what researchers might consider when attempting to design or report the highest quality or impact research. As a practical recommendation, researchers might consider the factors found to associate with more citations or higher quality in their field and critically evaluate which, if any, are relevant to their research. For example, given that longer articles tend to be more cited, a scholar might consider whether this fact might nudge them toward describing their research in more detail, conducting more substantial studies, or reporting multiple studies in one paper.

Several machine learning and regression analyses have shown that it is feasible to predict the long-term citation counts of papers to some extent, although with likely substantial disciplinary differences. It is difficult to quantify the disciplinary differences due to the differing methods, scopes and accuracy measures of the experiments reviewed. The most important inputs are probably journal properties, authorship team properties and field/topic properties, with early citation information being especially useful, for predicting long-term citation counts a few years after publication.

Machine learning has also been used to estimate the overall quality of articles from citation counts and bibliometric data, with the results suggesting that the methods are most accurate in medicine, the health and physical sciences but are inaccurate in the arts and humanities. Simpler methods have also been used to estimate the average quality of the articles of departments. Both approaches can give high overall correlations with peer review scores when averaged across departments but there are pragmatic reasons why even very high correlations may be insufficient to justify replacing peer review for important research evaluations.

Nevertheless, the levels of accuracy achievable for predicting long-term citations or article quality and the availability of science wide field-specific evidence of this (Thelwall, Kousha, Abdoli, Stuart, Makita, Wilson, & Cancellieri, 2023) suggest that it is now possible to use machine learning predictions of research quality for some fields to *support* peer review, especially where citation data alone currently performs this role. It is also possible to use machine learning predictions for formative national and departmental research evaluations that currently rely on citation-based indicators (e.g., government-commissioned

reports on national research performance, such as Elsevier, 2017). This could make the reports more accurate, albeit at the cost of greater complexity.

10.1 | Future research

In terms of future research, it is now possible to think about creating a shared dataset for the task of predicting long-term citation counts with the help of a scholarly database that is open to data sharing with the research community, such as Dimensions.ai, or free data sources, such as from CrossRef. The same dataset can be used for the task of finding properties that associate with long-term citation counts. This would allow many different methods to be evaluated on the same data, perhaps with agreed accuracy metrics, to help identify methods and inputs that are consistently useful. Such a dataset should be multidisciplinary with agreed splits into fields so that approaches that work differently between fields can be identified. This dataset would have the additional benefit of reproducibility. Consideration should be given to updating this annually, however, to keep pace with science, but encouraging authors to analyze older and newer slices of the data for comparability with earlier studies. This is not a perfect solution, however, since investigators may wish to collect their own additional data, such as altmetrics, or apply their own field classification schemes.

ACKNOWLEDGMENTS

This study was funded by Research England, Scottish Funding Council, Higher Education Funding Council for Wales, and Department for the Economy, Northern Ireland as part of the Future Research Assessment Programme (https://www.jisc.ac.uk/future-research-assessment-programme). 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

REFERENCES

Abramo, G., & D'Angelo, C. A. (2011). Evaluating research: From informed peer review to bibliometrics. *Scientometrics*, *87*(3), 499–514.

Abramo, G., & D'Angelo, C. A. (2015). The relationship between the number of authors of a publication, its citations and the impact factor of the publishing journal: Evidence from Italy. *Journal of Informetrics*, *9*(4), 746–761.

Abramo, G., D'Angelo, C. A., & Caprasecca, A. (2009). Allocative efficiency in public research funding: Can bibliometrics help? *Research Policy*, *38*(1), 206–215.

Abrishami, A., & Aliakbary, S. (2019). Predicting citation counts based on deep neural network learning techniques. *Journal of Informetrics*, *13*(2), 485–499.

Abt, H. A. (1984). Citations to single and multiauthored papers. *Publications of the Astronomical Society of the Pacific*, *96*(583), 746.

Ahlgren, P., Colliander, C., & Sjögårde, P. (2018). Exploring the relation between referencing practices and citation impact: A large-scale study based on Web of Science data. *Journal of the Association for Information Science and Technology*, *69*(5), 728–743.

Ahlgren, P., & Waltman, L. (2014). The correlation between citation-based and expert-based assessments of publication channels: SNIP and SJR vs. Norwegian quality assessments. *Journal of Informetrics*, *8*(4), 985–996.

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.

Aksnes, D. W. (2003). Characteristics of highly cited papers. *Research Evaluation*, *12*(3), 159–170.

Aksnes, D. W. (2006). Citation rates and perceptions of scientific contribution. *Journal of the American Society for Information Science and Technology*, *57*(2), 169–185.

Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *SAGE Open*, *9*(1), 2158244019829575.

Aksnes, D. W., & Taxt, R. E. (2004). Peer reviews and bibliometric indicators: A comparative study at a Norwegian university. *Research Evaluation*, *13*(1), 33–41.

Al-Janabi, S., Lim, L. W., & Aquili, L. (2021). Development of a tool to accurately predict UK REF funding allocation. *Scientometrics*, *126*(9), 8049–8062.

Allen, N., & Heath, O. (2013). Reputations and research quality in British political science: The importance of journal and publisher rankings in the 2008 RAE. *The British Journal of Politics and International Relations*, *15*(1), 147–162.

Ante, L. (2022). The relationship between readability and scientific impact: Evidence from emerging technology discourses. *Journal of Informetrics*, *16*(1), 101252.

Antonakis, J., Bastardoz, N., Liu, Y., & Schriesheim, C. A. (2014). What makes articles highly cited? *The Leadership Quarterly*, *25*(1), 152–179.

Asaad, M., Kallarackal, A. P., Meaike, J., Rajesh, A., De Azevedo, R. U., & Tran, N. V. (2020). Citation skew in plastic surgery journals: Does the journal impact factor predict individual article citation rate? *Aesthetic Surgery Journal*, *40*(10), 1136–1142.

Avkiran, N. (1997). Scientific collaboration in finance does not lead to better quality research. *Scientometrics*, *39*(2), 173–184.

Baccini, A., Barabesi, L., & De Nicolao, G. (2020). On the agreement between bibliometrics and peer review: Evidence from the Italian research assessment exercises. *PLoS One*, *15*(11), e0242520.

Balbuena, L. D. (2018). The UK research excellence framework and the Matthew effect: Insights from machine learning. *PLoS One*, *13*(11), e0207919.

Barker, C., & Pistrang, N. (2005). Quality criteria under methodological pluralism: Implications for conducting and evaluating research. *American Journal of Community Psychology*, *35*(3), 201–212.

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C. A., & Peracchi, F. (2015). Bibliometric evaluation vs. informed peer review: Evidence from Italy. *Research Policy*, *44*(2), 451–466.

BinMakhashen, G. M., & Al-Jamimi, H. A. (2022). Evaluation of machine learning to early detection of highly cited papers. In *2022 7th international conference on data science and machine learning applications (CDMA)* (pp. 1–6). IEEE.

Bishop, D. (2014). BishopBlog: An alternative to REF2014? Blogpost. http://deevybee.blogspot.nl/2013/01/an-alternative-to-ref2014.html

Blümel, C., & Schniedermann, A. (2020). Studying review articles in scientometrics and beyond: A research agenda. *Scientometrics*, *124*(1), 711–728.

Borgman, C. L., & Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology*, *36*(1), 1–53.

Bornmann, L., & Daniel, H. D. (2006). Selecting scientific excellence through committee peer review-a citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. *Scientometrics*, *68*(3), 427–440.

Bornmann, L., & Leydesdorff, L. (2015). Does quality and content matter for citedness? A comparison with para-textual factors and over time. *Journal of Informetrics*, *9*(3), 419–429.

Bornmann, L., Schier, H., Marx, W., & Daniel, H.-D. (2012). What factors determine citation counts of publications in chemistry besides their quality? *Journal of Informetrics*, *6*, 11–18.

Boyack, K. W., & Klavans, R. (2005). Predicting the importance of current papers. In *Proceedings of the 10th international conference of the International Society for Scientometrics and Informetrics* (Vol. 1, pp. 335–342). Karolinska University Press.

Buter, R. K., & van Raan, A. F. (2011). Non-alphanumeric characters in titles of scientific publications: An analysis of their occurrence and correlation with citation impact. *Journal of Informetrics*, *5*(4), 608–617.

Butler, L., & McAllister, I. (2009). Metrics or peer review? Evaluating the 2001 UK research assessment exercise in political science. *Political Studies Review*, *7*(1), 3–17.

Butler, L., & McAllister, I. (2011). Evaluating university research performance using metrics. *European Political Science*, *10*(1), 44–58.

Bordons, M., Aparicio, J., & Costas, R. (2013). Heterogeneity of collaboration and its relationship with research impact in a biomedical field. *Scientometrics*, *96*(2), 443–466.

Chi, P. S., & Glänzel, W. (2017). An empirical investigation of the associations among usage, scientific collaboration and citation impact. *Scientometrics*, *112*(1), 403–412.

Callaham, M., Wears, R. L., & Weber, E. (2002). Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *Jama*, *287*(21), 2847–2850.

Chen, J., & Zhang, C. (2015). Predicting citation counts of papers. In *2015 IEEE 14th international conference on Cognitive Informatics & Cognitive Computing (ICCI&CC)* (pp. 434–440). IEEE Press.

Colebunders, R., Kenyon, C., & Rousseau, R. (2014). Increase in numbers and proportions of review articles in tropical medicine, infectious diseases, and oncology. *Journal of the Association for Information Science and Technology*, *65*(1), 201–205.

Confraria, H., Godinho, M. M., & Wang, L. (2017). Determinants of citation impact: A comparative analysis of the global south versus the global north. *Research Policy*, *46*(1), 265–279.

Cummings, D., & Nassar, M. (2020). Structured citation trend prediction using graph neural networks. In *ICASSP 2020–2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 3897–3901). IEEE.

Cunil, O. M., González, L. O., Santomil, P. D., & Forteza, C. M. (2023). How to accomplish a highly cited paper in the tourism, leisure and hospitality field. *Journal of Business Research*, *157*, 113619.

Didegah, F. (2014). *Factors associating with the future citation impact of published articles: A statistical modelling approach*. (Doctoral dissertation). University of Wolverhampton. https://wlv.openrepository.com/handle/2436/322738

Didegah, F., & Thelwall, M. (2013a). Determinants of research citation impact in nanoscience and nanotechnology. *Journal of the American Society for Information Science and Technology*, *64*(55), 1055–1064.

Didegah, F., & Thelwall, M. (2013b). Which factors help authors produce the highest impact research? Collaboration, journal and document properties. *Journal of Informetrics*, *7*(4), 861–873.

Dowling, M., Hammami, H., & Zreik, O. (2018). Easy to read, easy to cite? *Economics Letters*, *173*, 100–103.

Elgendi, M. (2019). Characteristics of a highly cited article: A machine learning perspective. *IEEE Access*, *7*, 87977–87986.

Elsevier. (2017). An international comparison of the UK research base 2016. https://www.elsevier.com/research-intelligence?a=507321

Fairclough, R., & Thelwall, M. (2022). Questionnaires mentioned in academic research 1996-2019: Rapid increase but declining citation impact. *Learned Publishing*, *35*, 241–252.

Falagas, M. E., Zarkali, A., Karageorgopoulos, D. E., Bardakas, V., & Mavros, M. N. (2013). The impact of article length on the number of future citations: A bibliometric analysis of general medicine journals. *PLoS One*, *8*(2), e49476.

Fan, L., Guo, L., Wang, X., Xu, L., & Liu, F. (2022). Does the author's collaboration mode lead to papers' different citation impacts? An empirical analysis based on propensity score matching. *Journal of Informetrics*, *16*(4), 101350.

Fan, L., Wang, Y., Ding, S., & Qi, B. (2020). Productivity trends and citation impact of different institutional collaboration patterns at the research units' level. *Scientometrics*, *125*(2), 1179–1196.

Fiala, D., Král, P., & Dostal, M. (2021). Are papers asking questions cited more frequently in computer science? *Computers*, *10*(8), 96.

Figg, W. D., Dunn, L., Liewehr, D. J., Steinberg, S. M., Thurman, P. W., Barrett, J. C., & Birkinshaw, J. (2006). Scientific collaboration results in higher citation rates of published articles. *Pharmacotherapy*, *26*(6), 759–767.

Fox, C. W., Paine, C. T., & Sauterey, B. (2016). Citations increase with manuscript length, author number, and references cited in ecology journals. *Ecology and Evolution*, *6*(21), 7717–7726.

Franceschet, M., & Costantini, A. (2010). The effect of scholar collaboration on impact and quality of academic papers. *Journal of Informetrics*, *4*(4), 540–553.

Franceschet, M., & Costantini, A. (2011). The first Italian research assessment exercise: A bibliometric perspective. *Journal of Informetrics*, *5*(2), 275–291.

Fu, L., & Aliferis, C. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics*, *85*(1), 257–270.

Gazni, A. (2011). Are the abstracts of high impact articles more readable? Investigating the evidence from top research institutions in the world. *Journal of Information Science*, *37*(3), 273–281.

Gazni, A., & Didegah, F. (2011). Investigating different types of research collaboration and citation impact: A case study of Harvard University's publications. *Scientometrics*, *87*(2), 251–265.

Glänzel, W. (2002). Co-authorship patterns and trends in the sciences: A bibliometric study with implications for database indexing and search strategic 1980–1998. *Library Trends*, *50*(3), 461–473.

Gnewuch, M., & Wohlrabe, K. (2017). Title characteristics and citations in economics. *Scientometrics*, *110*(3), 1573–1578.

Graf-Vlachy, L., Graziotin, D., & Wagner, S. (2022). Text and team: what article metadata characteristics drive citations in Software Engineering? *Proceedings of the International Conference on Evaluation and Assessment in Software Engineering* (pp. 20–29).

Guo, F., Ma, C., Shi, Q., & Zong, Q. (2018). Succinct effect or informative effect: The relationship between title length and the number of citations. *Scientometrics*, *116*(3), 1531–1539.

Glänzel, W., Rinia, E. J., & Brocken, M. G. (1995). A bibliometric study of highly cited European physics papers in the 80s. *Research Evaluation*, *5*(2), 113–122.

Ha, T. (2022). An explainable artificial-intelligence-based approach to investigating factors that influence the citation of papers. *Technological Forecasting and Social Change*, *184*, 121974.

Habibzadeh, F., & Yadollahie, M. (2010). Are shorter article titles more attractive for citations? Crosssectional study of 22 scientific journals. *Croatian Medical Journal*, *51*(2), 165–170.

Haddow, G., & Genoni, P. (2010). Citation analysis and peer ranking of Australian social science journals. *Scientometrics*, *85*(2), 471–487.

Haddawy, P., Hassan, S. U., Asghar, A., & Amin, S. (2016). A comprehensive examination of the relation of three citation-based journal metrics to expert judgment of journal quality. *Journal of Informetrics*, *10*(1), 162–173.

Hafeez, D. M., Jalal, S., & Khosa, F. (2019). Bibliometric analysis of manuscript characteristics that influence citations: A comparison of six major psychiatry journals. *Journal of Psychiatric Research*, *108*, 90–94.

Hasan, S., & Breunig, R. (2021). Article length and citation outcomes. *Scientometrics*, *126*(9), 7583–7608.

Haslam, N., Ban, L., Kaufmann, L., Loughnan, S., Peters, K., Whelan, J., & Wilson, S. (2008). What makes an article influential? Predicting impact in social and personality psychology. *Scientometrics*, *76*(1), 169–185.

Haslam, N., & Koval, P. (2010). Predicting long-term citation impact of articles in social and personality psychology. *Psychological Reports*, *106*(3), 891–900.

Haustein, S., Costas, R., & Larivière, V. (2015). Characterizing social media metrics of scholarly papers: The effect of document properties and collaboration patterns. *PLoS One*, *10*(3), e0120495.

He, Z. L. (2009). International collaboration does not have greater epistemic authority. *Journal of the American Society for Information Science and Technology*, *60*(10), 2151–2164.

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/

Hodge, D. R., Victor, B. G., Grogan-Kaylor, A., & Perron, B. E. (2017). Disseminating high-impact social work scholarship: A longitudinal examination of 5-year citation count correlates. *Journal of the Society for Social Work and Research*, *8*(2), 211–231.

Hsu, J. W., & Huang, D. W. (2011). Correlation between impact and collaboration. *Scientometrics*, *86*(2), 317–324.

Hu, H., Wang, D., & Deng, S. (2021). Analysis of the scientific literature's abstract writing style and citations. *Online Information Review*, *45*(7), 1290–1305.

Hu, Y. H., Tai, C. T., Liu, K. E., & Cai, C. F. (2020). Identification of highly-cited papers using topic-model-based and bibliometric features: The consideration of keyword popularity. *Journal of Informetrics*, *14*(1), 101004.

Hussain, S., Almansouri, A., Allanqawi, L., Philteos, J., Wu, V., & Chan, Y. (2022). Does the journal impact factor predict individual article citation rate in otolaryngology journals? *Ear, Nose, & Throat Journal*, 1455613221119051.

Ibanez, A., Bielza, C., & Larranaga, P. (2013). Relationship among research collaboration, number of documents and number of citations: A case study in Spanish computer science production in 2000–2009. *Scientometrics*, *95*(2), 689–716.

Ibáñez, A., Larrañaga, P., & Bielza, C. (2009). Predicting citation count of Bioinformatics papers within four years of publication. *Bioinformatics*, *25*(24), 3303–3309.

Jacques, T. S., & Sebire, N. J. (2010). The impact of article titles on citation hits: An analysis of general and specialist medical journals. *JRSM Short Reports*, *1*(1), 1–5.

Jamali, H. R., & Nikzad, M. (2011). Article title type and its relation with the number of downloads and citations. *Scientometrics*, *88*(2), 653–661.

Jiang, F. K., & Hyland, K. (2022). 2Titles in research articles: Changes across time and discipline. *Learned Publishing*, *36*(2), 239–248.

Jump, P. (2015). Can the research excellence framework run on metrics? *Times Higher Education*. https://web.archive.org/web/20151013021233/https://www.timeshighereducation.com/can-the-research-excellence-framework-ref-run-on-metrics

Katz, J., & Hicks, D. (1997). How much is a collaboration worth? A calibrated bibliometric model. *Scientometrics*, *40*(3), 541–554.

Khor, K. A., & Yu, L. G. (2016). Influence of international co-authorship on the research citation impact of young universities. *Scientometrics*, *107*(3), 1095–1110.

Kousha, K., Thelwall, M., & 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.

Kumari, R., Uddin, A., Lee, B. H., & Choi, K. (2020). Analyzing the factors influencing the waiting time to first citation and long-term impact of publications. *Journal of Scientometric Research*, *9*(2), 127–135.

Lancho-Barrantes, B. S., Guerrero-Bote, V. P., & de Moya-Anegón, F. (2013). Citation increments between collaborating countries. *Scientometrics*, *94*(3), 817–831.

Lancho-Barrantes, B. S., Guerrero-Bote, V. P., & Moya-Anegón, F. (2010). What lies behind the averages and significance of citation indicators in different disciplines? *Journal of Information Science*, *36*(3), 371–382.

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.

Langham-Putrow, A., Bakker, C., & Riegelman, A. (2021). Is the open access citation advantage real? A systematic review of the citation of open access and subscription-based articles. *PLoS One*, *16*(6), e0253129.

Larivière, V., Gingras, Y., Sugimoto, C. R., & Tsou, A. (2015). Team size matters: Collaboration and scientific impact since 1900. *Journal of the Association for Information Science and Technology*, *66*(7), 1323–1332.

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature*, *504*(7479), 211–213.

Lee, D. (2020). Author-related factors predicting citation counts of conference papers: Focusing on computer and information science. *The Electronic Library*, *38*(3), 463–476.

Lee, P. S., West, J. D., & Howe, B. (2017). Viziometrics: Analyzing visual information in the scientific literature. *IEEE Transactions on Big Data*, *4*(1), 117–129.

Leimu, R., & Koricheva, J. (2005). What determines the citation frequency of ecological papers? *Trends in Ecology & Evolution*, *20*(1), 28–32.

Letchford, A., Preis, T., & Moat, H. S. (2016). The advantage of simple paper abstracts. *Journal of Informetrics*, *10*(1), 1–8.

Leydesdorff, L., Bornmann, L., & Wagner, C. S. (2019). The relative influences of government funding and international collaboration on citation impact. *Journal of the Association for Information Science and Technology*, *70*(2), 198–201.

Li, S., Zhao, W. X., Yin, E. J., & Wen, J. R. (2019). A neural citation count prediction model based on peer review text. *Proceedings of the 2019 conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 4914–4924).

Lokker, C., McKibbon, K. A., McKinlay, R. J., Wilczynski, N. L., & Haynes, R. B. (2008). Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: Retrospective cohort study. *BMJ*, *336*(7645), 655–657.

Löwe, B. (2022). Measuring the agreement of mathematical peer reviewers. *Axiomathes*, *32*, 1205–1219.

Lyu, P. H., & Wolfram, D. (2018). Do longer articles gather more citations? Article length and scholarly impact among top biomedical journals. *Proceedings of the Association for Information Science and Technology*, *55*(1), 319–326.

Ma, A., Liu, Y., Xu, X., & Dong, T. (2021). A deep-learning based citation count prediction model with paper metadata semantic features. *Scientometrics*, *126*(8), 6803–6823.

Mahdi, S., D'Este, P., & Neely, A. (2008). *Citation counts: Are they good predictors of RAE scores?* Advanced Institute of Management Research. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1154053

Mammola, S., Fontaneto, D., Martínez, A., & Chichorro, F. (2021). Impact of the reference list features on the number of citations. *Scientometrics*, *126*(1), 785–799.

Mammola, S., Piano, E., Doretto, A., Caprio, E., & Chamberlain, D. (2022). Measuring the influence of non-scientific features on citations. *Scientometrics*, *127*(7), 4123–4137.

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & Delgado López-Cózar, E. (2021). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A multidisciplinary comparison of coverage via citations. *Scientometrics*, *126*(1), 871–906. https://doi.org/10.1007/s11192-020-03690-4

Medoff, M. H. (2003). Collaboration and the quality of economics research. *Labour Economics*, *10*(5), 597–608.

Mingers, J., & Xu, F. (2010). The drivers of citations in management science journals. *European Journal of Operational Research*, *205*(2), 422–430.

Miranda, R., & Garcia-Carpintero, E. (2018). Overcitation and overrepresentation of review papers in the most cited papers. *Journal of Informetrics*, *12*(4), 1015–1030.

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics*, *106*, 213–228.

Mryglod, O., Kenna, R., Holovatch, Y., & Berche, B. (2013). Comparison of a citation-based indicator and peer review for absolute and specific measures of research-group excellence. *Scientometrics*, *97*(3), 767–777.

Mryglod, O., Kenna, R., Holovatch, Y., & Berche, B. (2015a). Predicting results of the Research Excellence Framework using departmental h-index. *Scientometrics*, *102*(3), 2165–2180.

Mryglod, O., Kenna, R., Holovatch, Y., & Berche, B. (2015b). Predicting results of the research excellence framework using departmental h-index: Revisited. *Scientometrics*, *104*, 1013–1017.

Narin, F., Stevens, K., & Whitlow, E. S. (1991). Scientific co-operation in Europe and the citation of multinationally authored papers. *Scientometrics*, *21*(3), 313–323.

Norris, M., & Oppenheim, C. (2003). Citation counts and the research assessment exercise V: Archaeology and the 2001 RAE. *Journal of Documentation*, *59*(6), 709–730.

Norris, M., & Oppenheim, C. (2010). Peer review and the h-index: Two studies. *Journal of Informetrics*, *4*(3), 221–232.

Nomaler, Ö., Frenken, K., & Heimeriks, G. (2013). Do more distant collaborations have more citation impact? *Journal of Informetrics*, *7*(4), 966–971.

Onodera, N., & Yoshikane, F. (2015). Factors affecting citation rates of research articles. *Journal of the Association for Information Science and Technology*, *66*(4), 739–764.

Oppenheim, C. (1995). The correlation between citation counts and the 1992 research assessment exercise ratings for British Library and Information Science university departments. *Journal of Documentation*, *51*(1), 18–27.

Oppenheim, C. (1997). The correlation between citation counts and the 1992 research assessment exercise ratings for British research in genetics, anatomy and archaeology. *Journal of Documentation*, *53*(5), 477–487.

Oppenheim, C., & Summers, M. (2008). Citation counts and the research assessment exercise, part VI: Unit of assessment 67 (music). *Information Research*, *13*(2). http://InformationR.net/ir/13-2/paper342.html

Paiva, C. E., Lima, J. P. S. N., & Paiva, B. S. R. (2012). Articles with short titles describing the results are cited more often. *Clinics*, *67*, 509–513.

Peng, T. Q., & Zhu, J. J. (2012). Where you publish matters most: A multilevel analysis of factors affecting citations of internet studies. *Journal of the American Society for Information Science and Technology*, *63*(9), 1789–1803.

Peters, H. P., & van Raan, A. F. (1994). On determinants of citation scores: A case study in chemical engineering. *Journal of the American Society for Information Science*, *45*(1), 39–49.

Pride, D., & Knoth, P. (2018). Peer review and citation data in predicting university rankings, a large-scale analysis. In

International conference on theory and practice of digital libraries (pp. 195–207). Springer.

Puuska, H. M., Muhonen, R., & Leino, Y. (2014). International and domestic co-publishing and their citation impact in different disciplines. *Scientometrics*, *98*(2), 823–839.

Qian, Y., Rong, W., Jiang, N., Tang, J., & Xiong, Z. (2017). Citation regression analysis of computer science publications in different ranking categories and subfields. *Scientometrics*, *110*(3), 1351–1374.

Reale, E., Barbara, A., & Costantini, A. (2007). Peer review for the evaluation of academic research: Lessons from the Italian experience. *Research Evaluation*, *16*(3), 216–228.

Rinia, E. J., van Leeuwen, T. N., Van Vuren, H. G., & Van Raan, A. F. (1998). Comparative analysis of a set of bibliometric indicators and central peer review criteria: Evaluation of condensed matter physics in the Netherlands. *Research Policy*, *27*(1), 95–107.

Robson, B. J., & Mousquès, A. (2016). Can we predict citation counts of environmental modelling papers? Fourteen bibliographic and categorical variables predict less than 30% of the variability in citation counts. *Environmental Modelling & Software*, *75*, 94–104.

Rodríguez-Navarro, A., & Brito, R. (2020). Like-for-like bibliometric substitutes for peer review: Advantages and limits of indicators calculated from the ep index. *Research Evaluation*, *29*(2), 215–230.

Roldan-Valadez, E., & Rios, C. (2015). Alternative bibliometrics from impact factor improved the esteem of a journal in a 2-year-ahead annual-citation calculation: Multivariate analysis of gastroenterology and hepatology journals. *European Journal of Gastroenterology & Hepatology*, *27*(2), 115–122.

Ronda-Pupo, G. A. (2017). The effect of document types and sizes on the scaling relationship between citations and co-authorship patterns in management journals. *Scientometrics*, *110*(3), 1191–1207.

Rostami, F., Mohammadpoorasl, A., & Hajizadeh, M. (2014). The effect of characteristics of title on citation rates of articles. *Scientometrics*, *98*(3), 2007–2010.

Rousseau, R. (1992). Why am I not cited or, why are multi-authored papers more cited than others? *Journal of Documentation*, *48*(1), 79–80.

Rousseau, R., & Ding, J. (2016). Does international collaboration yield a higher citation potential for US scientists publishing in highly visible interdisciplinary journals? *Journal of the Association for Information Science and Technology*, *67*(4), 1009–1013.

Royle, P., Kandala, N. B., Barnard, K., & Waugh, N. (2013). Bibliometrics of systematic reviews: Analysis of citation rates and journal impact factors. *Systematic Reviews*, *2*(1), 1–11.

Ruan, X., Zhu, Y., Li, J., & Cheng, Y. (2020). Predicting the citation counts of individual papers via a BP neural network. *Journal of Informetrics*, *14*(3), 101039.

Saeed, A. U., Afzal, M. T., Latif, A., & Tochtermann, K. (2008). Citation rank prediction based on bookmark counts: Exploratory case study of WWW06 papers. In *2008 IEEE international multitopic conference* (pp. 392–397). IEEE Press.

Satish, N. G. (2021). International collaboration and high citation impact – A case analysis of immunology. *Annals of Library and Information Studies (ALIS)*, *68*(4), 366–375.

Sebo, P., & Clair, C. (2023). Gender inequalities in citations of articles published in high-impact general medical journals: A cross-sectional study. *Journal of General Internal Medicine*, *38*(3), 661–666. https://doi.org/10.1007/s11606-022-07717-9

Seng, L. B., & Willett, P. (1995). The citedness of publications by United Kingdom library schools. *Journal of Information Science*, *21*(1), 68–71.

Shen, H., Xie, J., Li, J., & Cheng, Y. (2021). The correlation between scientific collaboration and citation count at the paper level: A meta-analysis. *Scientometrics*, *126*(4), 3443–3470.

Sienkiewicz, J., & Altmann, E. G. (2016). Impact of lexical and sentiment factors on the popularity of scientific papers. *Royal Society Open Science*, *3*(6), 160140.

Sin, S. C. J. (2011). International coauthorship and citation impact: A bibliometric study of six LIS journals, 1980–2008. *Journal of the American Society for Information Science and Technology*, *62*(9), 1770–1783.

Singh, V. K., Singh, P., Karmakar, M., Leta, J., & Mayr, P. (2021). The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. *Scientometrics*, *126*, 5113–5142.

Sjögårde, P., & Didegah, F. (2022). The association between topic growth and citation impact of research publications. *Scientometrics*, *127*(4), 1903–1921.

Slyder, J. B., Stein, B. R., Sams, B. S., Walker, D. M., Jacob Beale, B., Feldhaus, J. J., & Copenheaver, C. A. (2011). Citation pattern and lifespan: A comparison of discipline, institution, and individual. *Scientometrics*, *89*(3), 955–966.

Small, H. (2018). Characterizing highly cited method and non-method papers using citation contexts: The role of uncertainty. *Journal of Informetrics*, *12*(2), 461–480.

Smith, A., & Eysenck, M. (2002). The correlation between RAE ratings and citation counts in psychology. June 2002. http://cogprints.org/2749/1/citations.pdf

Smith, M. J., Weinberger, C., Bruna, E. M., & Allesina, S. (2014). The scientific impact of nations: Journal placement and citation performance. *PLoS One*, *9*(10), e109195.

Sohrabi, B., & Iraj, H. (2017). The effect of keyword repetition in abstract and keyword frequency per journal in predicting citation counts. *Scientometrics*, *110*(1), 243–251.

Sonnenwald, D. H. (2007). Scientific collaboration. *Annual Review of Information Science and Technology*, *41*(1), 643–681.

Sooryamoorthy, R. (2009). Do types of collaboration change citation? Collaboration and citation patterns of south African science publications. *Scientometrics*, *81*(1), 177–193.

Stremersch, S., Verniers, I., & Verhoef, P. C. (2007). The quest for citations: Drivers of article impact. *Journal of Marketing*, *71*(3), 171–193.

Subotic, S., & Mukherjee, B. (2014). Short and amusing: The relationship between title characteristics, downloads, and citations in psychology articles. *Journal of Information Science*, *40*(1), 115–124.

Sud, P., & Thelwall, M. (2016). Not all international collaboration is beneficial: The Mendeley readership and citation impact of biochemical research collaboration. *Journal of the Association for Information Science and Technology*, *67*(8), 1849–1857.

Tahamtan, I., Safipour Afshar, A., & Ahamdzadeh, K. (2016). Factors affecting number of citations: A comprehensive review of the literature. *Scientometrics*, *107*, 1195–1225.

Taylor, J. (2011). The assessment of research quality in UK universities: Peer review or metrics? *British Journal of Management*, *22*(2), 202–217.

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. (2018). Do females create higher impact research? Scopus citations and Mendeley readers for articles from five countries. *Journal of Informetrics*, *12*(4), 1031–1041. https://doi.org/10.1016/j.joi.2018.08.005

Thelwall, M. (2020a). Female citation impact superiority 1996-2018 in six out of seven English-speaking nations. *Journal of the Association for Information Science and Technology*, *71*(8), 979–990. https://doi.org/10.1002/asi.24316

Thelwall, M. (2020b). Gender differences in citation impact for 27 fields and 6 English speaking countries 1996-2014. *Quantitative Science Studies*, *1*(2), 599–617.

Thelwall, M. (2022). Can the quality of published academic journal articles be assessed with machine learning? *Quantitative Science Studies*, *3*(1), 208–226.

Thelwall, M. (2023). Are successful co-authors more important than first authors for publishing academic journal articles? *Scientometrics*, *128*(4), 2211–2232. https://doi.org/10.1007/s11192-023-04663-z

Thelwall, M., Kousha, K., Abdoli, M., Stuart, E., Makita, M., Font-Julián, C., Wilson, P., & Levitt, J. (2023). Is research funding always beneficial? A cross-disciplinary analysis of UK research 2014-20. *Quantitative Science Studies*, 1–34. https://doi.org/10.1162/qss_a_00254

Thelwall, M., Kousha, K., Abdoli, M., Stuart, E., Makita, M., Wilson, P., & Cancellieri, M. (2023). Predicting article quality scores with machine learning: The UK research excellence framework. *Quantitative Science Studies*, 1–27. https://doi.org/10.1162/qss_a_00258

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

Thelwall, M., Kousha, K., Abdoli, M., Stuart, E., Makita, M., Wilson, P., & Levitt, J. (2022b). Do bibliometrics introduce gender, institutional or interdisciplinary biases into research evaluations? *Research Policy*, *52*(8). https://doi.org/10.1016/j.respol.2023.104829

Thelwall, M., Kousha, K., Abdoli, M., Stuart, E., Makita, M., Wilson, P., & Levitt, J. (2022c). In which fields do higher impact journals publish higher quality articles? *Scientometrics*, *128*, 3915–3933. https://doi.org/10.1007/s11192-023-04735-0

Thelwall, M., Kousha, K., Abdoli, M., Stuart, E., Makita, M., Wilson, P., & Levitt, J. (2022d). Are internationally co-authored journal articles better quality? The UK case 2014-2020. *arXiv*, arXiv:2212.05417.

Thelwall, M., Kousha, K., Abdoli, M., Stuart, E., Makita, M., Wilson, P., & Levitt, J. (2023a). In which fields are citations indicators of research quality? *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi.24767

Thelwall, M., Kousha, K., Abdoli, M., Stuart, E., Makita, M., Wilson, P., & Levitt, J. (2023d). Why are co-authored academic articles more cited: Higher quality or larger audience? *Journal of the Association for Information Science and Technology*, *74*, 791–810. https://doi.org/10.1002/asi.24755

Thelwall, M., & Maflahi, N. (2020). Academic collaboration rates and citation associations vary substantially between countries and fields. *Journal of the Association for Information Science and Technology*, *71*(8), 968–978.

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., & Nevill, T. (2021). Is research with qualitative data more prevalent and impactful now? Interviews, case studies, focus groups and ethnographies. *Library & Information Science Research*, *43*(2), 101094. https://doi.org/10.1016/j.lisr.2021.101094

Thelwall, M., & Sud, P. (2014). No citation advantage for monograph-based collaborations? *Journal of Informetrics*, *8*(1), 276–283.

Thelwall, M., & Sud, P. (2022). Scopus 1900-2020: Growth in articles, abstracts, countries, fields, and journals. *Quantitative Science Studies*, *3*(1), 37–50.

Thelwall, M., & Wilson, P. (2016). Does research with statistics have more impact? The citation rank advantage of structural equation modelling. *Journal of the Association for Information Science and Technology*, *67*(5), 1233–1244. https://doi.org/10.1002/asi.23474

Thomas, P., & Watkins, D. (1998). Institutional research rankings via bibliometric analysis and direct peer review: A comparative case study with policy implications. *Scientometrics*, *41*(3), 335–355.

Tourish, D. (2020). The triumph of nonsense in management studies. *Academy of Management Learning & Education*, *19*(1), 99–109.

Traag, V. A., & Waltman, L. (2019). Systematic analysis of agreement between metrics and peer review in the UK REF. Palgrave. *Communications*, *5*(1), 29.

Trowler, P. (2014). Academic tribes and territories: The theoretical trajectory. *Österreichische Zeitschrift für Geschichtswissenschaften*, *25*(3), 17–26.

Urlings, M. J., Duyx, B., Swaen, G. M., Bouter, L. M., & Zeegers, M. P. (2021). Citation bias and other determinants of citation in biomedical research: Findings from six citation networks. *Journal of Clinical Epidemiology*, *132*, 71–78.

van Dalen, H. P., & Henkens, K. N. (2005). Signals in science – On the importance of signaling in gaining attention in science. *Scientometrics*, *64*(2), 209–233.

Van Raan, A. (1998). The influence of international collaboration on the impact of research results: Some simple mathematical considerations concerning the role of self-citations. *Scientometrics*, *42*(3), 423–428.

Van Raan, A. F. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, *67*(3), 491–502.

Van Wesel, M., Wyatt, S., & ten Haaf, J. (2014). What a difference a colon makes: How superficial factors influence subsequent citation. *Scientometrics*, *98*(3), 1601–1615.

Vanclay, J. K. (2013). Factors affecting citation rates in environmental science. *Journal of Informetrics*, *7*(2), 265–271.

Vieira, E. S., & Gomes, J. A. N. F. (2010). Citation to scientific articles: Its distribution and dependence on the article features. *Journal of Informetrics*, *4*(1), 1–13.

Wagner, C. S., Whetsell, T. A., & Mukherjee, S. (2019). International research collaboration: Novelty, conventionality, and atypicality in knowledge recombination. *Research Policy*, *48*(5), 1260–1270.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C., Tijssen, R. J., van Eck, N. J., & Wouters, P. (2012). The Leiden ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, *63*(12), 2419–2432.

Wang, L., Thijs, B., & Glänzel, W. (2015). Characteristics of international collaboration in sport sciences publications and its influence on citation impact. *Scientometrics*, *105*(2), 843–862.

Wang, M., Jiao, S., Zhang, J., Zhang, X., & Zhu, N. (2020). Identification high influential articles by considering the topic characteristics of articles. *IEEE Access*, *8*, 107887–107899.

Wang, M., Yu, G., An, S., & Yu, D. (2012). Discovery of factors influencing citation impact based on a soft fuzzy rough set model. *Scientometrics*, *93*(3), 635–644.

Wang, M., Yu, G., & Yu, D. (2011). Mining typical features for highly cited papers. *Scientometrics*, *87*(3), 695–706.

Wang, S., Liu, X., & Zhou, J. (2022). Readability is decreasing in language and linguistics. *Scientometrics*, *127*(8), 4697–4729.

Weale, A. R., Bailey, M., & Lear, P. A. (2004). The level of non-citation of articles within a journal as a measure of quality: A comparison to the impact factor. *BMC Medical Research Methodology*, *4*(1), 1–8.

Weinberger, C. J., Evans, J. A., & Allesina, S. (2015). Ten simple (empirical) rules for writing science. *PLoS Computational Biology*, *11*(4), e1004205.

Whitley, R. (2000). *The intellectual and social organization of the sciences* (2nd ed.). Oxford University Press on Demand.

Willis, D. L., Bahler, C. D., Neuberger, M. M., & Dahm, P. (2011). Predictors of citations in the urological literature. *BJU International*, *107*(12), 1876–1880.

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Kain, R., Kerridge, S., Thelwall, M., Tinkler, J., Viney, I., Wouters, P., Hill, J., & Johnson, B. (2015). The metric tide: Report of the independent review of the role of metrics in research assessment and management. https://doi.org/10.13140/RG.2.1.4929.1363

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*(5827), 1036–1039.

Xiao, L., & Jiang, W. (2020). Correlation between references and citations in artificial intelligence: A preliminary study. *Proceedings of the Association for Information Science and Technology*, *57*(1), e403.

Xie, J., Gong, K., Cheng, Y., & Ke, Q. (2019). The correlation between paper length and citations: A meta-analysis. *Scientometrics*, *118*(3), 763–786.

Xu, J., Li, M., Jiang, J., Ge, B., & Cai, M. (2019). Early prediction of scientific impact based on multi-bibliographic features and convolutional neural network. *IEEE Access*, *7*, 92248–92258.

Yu, T., Yu, G., Li, P. Y., & Wang, L. (2014). Citation impact prediction for scientific papers using stepwise regression analysis. *Scientometrics*, *101*(2), 1233–1252.

Zhao, Q., & Feng, X. (2022). Utilizing citation network structure to predict paper citation counts: A deep learning approach. *Journal of Informetrics*, *16*(1), 101235.

How to cite this article: Kousha, K., & Thelwall, M. (2023). Factors associating with or predicting more cited or higher quality journal articles: An Annual Review of Information Science and Technology (ARIST) paper. *Journal of the Association for Information Science and Technology*, 1–30. https://doi.org/10.1002/asi.24810