University of Dundee

In which fields are citations indicators of research quality?

Thelwall, Mike; Kousha, Kayvan; Abdoli, Mahshid; Stuart, Emma; Makita, Meiko; Wilson, Paul

Document Version
Early version, also known as pre-print

Link to publication in Discovery Research Portal

In which fields are citations indicators of research quality?

Mike Thelwall
Statistical Cybermetrics and Research Evaluation Group, University of Wolverhampton, UK.
https://orcid.org/0000-0001-6065-205X m.thelwall@wlv.ac.uk

Kayvan Kousha
Statistical Cybermetrics and Research Evaluation Group, University of Wolverhampton, UK.
https://orcid.org/0000-0003-4827-971X k.kousha@wlv.ac.uk

Emma Stuart
Statistical Cybermetrics and Research Evaluation Group, University of Wolverhampton, UK.
https://orcid.org/0000-0003-4807-7659 emma.stuart@wlv.ac.uk

Meiko Makita
Statistical Cybermetrics and Research Evaluation Group, University of Wolverhampton, UK.
https://orcid.org/0000-0002-2284-0161 meikomakita@wlv.ac.uk

Mahshid Abdoli
Statistical Cybermetrics and Research Evaluation Group, University of Wolverhampton, UK.
https://orcid.org/0000-0001-9251-5391 m.abdoli@wlv.ac.uk

Paul Wilson
Statistical Cybermetrics and Research Evaluation Group, University of Wolverhampton, UK.
https://orcid.org/0000-0002-1265-543X pauljwilson@wlv.ac.uk

Jonathan Levitt
Statistical Cybermetrics and Research Evaluation Group, University of Wolverhampton, UK.
https://orcid.org/0000-0002-4386-3813 j.m.levitt@wlv.ac.uk

Citation counts are widely used as indicators of research quality to support or replace human peer review and for lists of top cited papers, researchers, and institutions. Nevertheless, the extent to which citation counts reflect research quality is not well understood. We report the largest-scale evaluation of the relationship between research quality and citation counts, correlating them for 87,739 journal articles in 34 field-based Units of Assessment (UoAs) from the UK. We show that the two correlate positively in all academic fields examined, from very weak (0.1) to strong (0.5). The highest correlations are in health, life sciences and physical sciences and the lowest are in the arts and humanities. The patterns are similar for the field classification schemes of Scopus and Dimensions.ai. We also show that there is no citation threshold in any field beyond which all articles are excellent quality, so lists of top cited articles are not definitive collections of excellence. Moreover, log transformed citation counts have a close to linear relationship with UK research quality ranked scores that is shallow in some fields but steep in others. In conclusion, whilst appropriately field normalised citations associate positively with research quality in all fields, they never perfectly reflect it, even at very high values.
Keywords: Research evaluation; Citation analysis; Research quality; Research Excellence Framework; REF2021; Scopus.

1   Introduction

Citations are widely used as formal or informal indicators of research value. In Italy, for example, articles with enough citations and published in a journal with enough citations per paper used to be automatically classified as high quality in the national research assessment exercise (Abramo & D'Angelo, 2016). Citations also inform peer review exercises (REF2021, 2020), organisations such as Clarivate celebrate highly cited articles and researchers, and university league tables often include a citation-based component (Waltman, et al., 2012). Citation-based Journal Impact Factors (JIFs) and variants are frequently used for recognition and reward too (McKiernan et al., 2019). Nevertheless, the value of citations for research assessment is contested and controversial. Research evaluation guidelines caution against overreliance on research metrics (Hicks et al., 2015; Wilsdon et al., 2015), and there are strong arguments against using citations for aspects of research assessment because they do not always reflect impact and ignore some article influences (e.g., MacRoberts & MacRoberts, 2010). Many organisations have also signed the San Francisco Declaration of Research Assessment (DORA), which argues against reliance on journal impact factors (sfdora.org). In the UK, most national research evaluation avoids all bibliometrics (REF2021, 2020). Despite the widespread use of citation counts and their controversial nature, it is still not clear which fields they are appropriate for and how they relate to research quality. This article provides evidence using a large sample of journal articles with expert peer review quality scores from the UK.

Research quality is a vague overall concept but is usually thought of in terms of methodological rigour, novelty/originality, and impact on science or society (Langfeldt et al., 2020). The way in which these three aspects is tested varies substantially between fields. For example, to be judged 'world leading', research might need to be "a primary or essential point of reference" in the arts and humanities or impress in terms of "the scale, challenge and logistical difficulty posed by the research" for health-related studies (REF2021, 2019). Of the three aspects of research quality, citations best reflect impact on science (Aksnes et al., 2019), so citation counts may undervalue research that is particularly strong for rigour, novelty or impact on society. Since there may be a tendency for research to be simultaneously strong or weak in all aspects of quality, and this may vary between fields, it is not clear whether citation counts are reasonable indicators of overall quality in any or all fields.

A core theoretical basis for using citation counts as an indicator of research quality, or at least its scholarly impact dimension, is that citations serve to acknowledge the relevant prior work of other scholars. Thus, counting the citations to an article might give a measure of how often it has proven useful (Merton, 1973). There are several arguments against this, however. First, there are many reasons to cite prior work, including for background context, to refute, and to show improvement without necessarily drawing upon the cited work (Lyu et al., 2021). Second, humanities fields are non-hierarchical and there is much less need to build on the work of other scholars (Whitley, 2000). Third, there are many factors that influence the choice of citations, such as the tendency to cite known scholars, friends, or editors, biasing the counts (Borgman & Furner, 2002; Vinkler, 1987). Fourth, non-journal outputs (e.g., reports, books, art) are important in some fields (Hicks, 2004) but are often largely or fully excluded from citation databases. A statistical response to criticisms like these is to accept that there are reasons for citing that do not reflect impact or that reflect little impact, but that when citations are aggregated on a sufficiently large scale then the "imperfections" may tend to average out. This would allow indicators based on average citations to have some value, even if they do not work well for individual journal articles (van Raan, 2004). Since the

amount of bias and the amount of "signal to noise" in citations is unknown, the task of identifying the contexts, such as fields and years, in which it is appropriate to use citation-based indicators is essentially a statistical one: assessing if and when citation counts correlate to a sufficient degree with article quality.

Many studies have compared public aggregate evidence of research quality with average citations for collections of outputs, with mixed results. Rankings of UK departments based on average peer review scores for their outputs have been compared to average citation-based rankings, with correlations being very strong (rho=0.9) for psychology (Smith & Eysenck, 2002) library and information science (rho=0.8) (Norris & Oppenheim, 2003; Oppenheim, 1995), Archaeology (rho=0.7), Genetics (rho=0.7), and Anatomy (rho=0.5) (Seng & Willett, 1995), and Music (rho=0.8) (Oppenheim & Summers, 2008) and political science (partial correlation: 0.5) (Butler & McAllister, 2009). A larger scale study found strong associations between average citations and average REF scores for journal articles in life and health sciences (except nursing), business and economics, but weak associations in the social sciences (Mahdi, D'Este & Neely, 2008). High correlations (0.7-0.8) between departmental REF2014 output rankings and median citations per paper have also been found for ten UoAs (Pride & Knoth, 2018).

Outside the UK, an investigation into 12,000 Italian research articles correlated institutional average peer review scores with institutional average citations per paper in ten fields. There were strong correlations in most, including Physics (rho=0.8), Earth Sciences (0.8), Biology (0.7), and Chemistry (0.6) (Franceschet & Costantini, 2011), but weaker correlations have been found with a different method for Italy (except medicine, 0.5: Abramo et al, 2011; Baccini & De Nicolao, 2016). High correlations have also been obtained for the Netherlands (Rinia et al., 1998; van Raan, 2006). From a related perspective, panel ratings had weak correlations with citation-based indicators for research groups within an institution in Norway (Aksnes & Taxt, 2004). These studies give little information about the strength of article level correlations within fields, however, because correlations increase in magnitude when they are aggregated, with the degree of increase depending on the size of the aggregation units. Thus, it is not possible to draw conclusions about article-level correlations from institution-level correlations.

A few studies have directly investigated the extent to which citation counts correlate with research quality for journal articles. The largest scale study investigated peer review scores for about 25,000 journal articles published in 2008 with citation-based indicators in 36 UK Research Excellence Framework (REF) Units of Assessment (UoAs) and reported weaker results for articles from 2013 (HEFCE, 2015). Overall, REF peer review scores for individual articles significantly and positively correlated (0.3) with Elsevier's field normalised citation impact metric Source-Normalised Impact per Paper (SNIP), Field-Weighted Citation Impact (0.3), and citation counts (0.2). There were large disciplinary differences within this overall figure, with the strongest correlations between citation counts and REF scores in Clinical Medicine (rho=0. 7), Chemistry (0.6), Physics (0.6) and Biological Sciences (0.6). Correlations in most social sciences, arts and humanities were typically below 0.3 (HEFCE, 2015). This is probably the best available evidence of the relationship between research quality and citation counts at a relatively fine-grained level, but the results did not report confidence intervals and only used a single year of data, limiting the conclusions that can be drawn. Moreover, UK UoAs are unique aggregation units that do not easily map onto other field classification schemes, which also limits the generalisability of the results.

Given the lack of conclusive evidence about the relationship between research quality and citations in the different academic fields, this article addresses the following research questions.

- RQ1: In which fields do more cited standard journal articles (excluding reviews) tend to be higher quality?
- RQ2: Does the answer to the above depend on the field classification scheme used?
- RQ3: Is there a citation threshold in any field, above which all research is high quality?
- RQ4: What is the overall shape of the relationship between citations and research quality?

2 Methods

2.1 Data

The data analysed in this paper is a subset of the journal articles submitted to UK REF2021. For this, each active higher education researcher in the UK had to submit between 1 and 5 outputs first published between 2014 and 2020, with an average of 2.5 outputs per full time equivalent member of staff. These outputs were submitted to one of 34 Units of Assessment (UoAs) and were then individually evaluated by at least two UoA subject specialists and awarded a quality score for "originality, significance and rigour" of 1* "recognised nationally", 2* "recognised internationally", 3* "internationally excellent", or 4* "world-leading". Outputs judged ineligible or below national quality were scored 0 instead (REF2021, 2020).

The subject specialists evaluating outputs were mainly senior UK researchers, although some international researchers and industry figures also participated. All types of research output could be submitted but only journal articles are analysed here. Review-type outputs are ineligible for the REF, so all articles report original research. Each author of a paper is entitled to submit it, but two authors from the same institution are usually not allowed to submit the same output.

Provisional REF2021 scores were supplied in March 2022 for 148,977 journal articles, which is an almost complete set except those from the University of Wolverhampton for confidentiality. The articles were matched with journal articles in Scopus with a recorded date between 2014 to 2018 that were downloaded in January 2021, to coincide with the date when the REF2021 evaluation was scheduled to start (although it was delayed by Covid-19). The matching was by DOI (99%) or by title, year and journal manually checked (1%). Articles from after 2018 were excluded to give a citation window of at least two years for analysis, and the 318 articles scoring 0 were removed because these had often not been evaluated for quality.

The citation counts for the journal articles were transformed into field and year normalised scores to allow different fields and years to be merged. To calculate Normalised Log-transformed Citation Scores (NLCS) (Thelwall, 2017), all Scopus articles 2014-18 were first log-transformed with $\ln(1+x)$ to reduce skewing and prevent the normalisation calculations from being heavily influenced by individual highly cited articles. After this, in each of the 326 Scopus narrow fields and years, the average log-transformed citation count was calculated. Finally, the log-transformed citation count $\ln(1+x)$ of each matching REF2021 article was divided by the average just calculated for the field and year in which it was published. Articles in multiple fields were divided instead by the average of the averages of the fields containing them. An NLCS for an article of 1 always equates to world average citation count for its Scopus-indexed field(s) and year. Values higher than 1 always mean more cited than average

for the publishing field(s) and year. The NLCS values were grouped into fields for analysis and compared with the provisional REF2021 quality scores for the same articles.

REF2021 organises the evaluation in 34 Units of Assessment grouped into four Main Panels, but there are other ways of grouping research into fields and so two alternative categorisation schemes were also used: The article-based Artificial Intelligence (AI) scheme of Dimensions.ai and the mainly journal-based scheme of Scopus. For Dimensions, each REF2021 article with a DOI was matched against Dimensions records with an API DOI search. The top-level Field Of Research (FOR) codes reported by Dimensions for each matching article were saved for the matching record. For Scopus, the top-level broad fields were used, as recorded by the Scopus API.

After all data processing, there were 87,739 journal articles across the 34 UoAs and 83,327 across the 4 main panels. The reason for the difference is that duplicate articles were eliminated within groups so that each UoA or Main Panel dataset includes no duplicates. There were many duplicates between UoAs (i.e., the same article submitted to multiple UoAs by different authors), so more duplicates were removed when forming main panel groups than UoA groups. For Scopus, a total of 144,207 articles were analysed and for Dimensions 99,661 articles were analysed. In both cases, the figures include duplicates between but not within fields or UoAs.

2.2   Analysis

The REF and NLCS scores were compared primarily through Spearman correlations. Although Pearson correlations would have been reasonable, given the log transformation to reduce skewing, Spearman is a conservative option and is appropriate given that REF scores occur on a non-scalar four-point qualitative system. Confidence intervals for the correlations were calculated using standard Fisher (1915) transformations.

For RQ3 and RQ4, the articles were bucketed into groups of at least 25 for analysis. This had two purposes. First, the scores for individual REF outputs are confidential and so individual values cannot be shown. Second, a reasonable sample size is necessary to differentiate between coincidence and trend. For example, to address RQ3, if the most cited article in UoA 1 had a 4* quality score, it would not be reasonable to give a positive answer on the basis of the citation count of that article. The bucket size of 25 seems like a reasonable compromise between too fine grained and too broad. This relatively arbitrary bucket size is a limitation, however.

3   Results

3.1   RQ1, RQ2: overall relationship between citations and research quality

Articles with more citations tend to be higher quality in all fields of science, whether using the REF (Figure 1), Dimensions (Figure 2) or Scopus (Figure 3) classification schemes. Perhaps surprisingly, there are statistically significant positive correlations even in most arts and humanities fields, including UoA 33: Music, Drama, Dance and Performing Arts (Figure 1), Studies in Creative Arts and Writing (Figure 2) and Arts and Humanities (Figure 3). Increasing the citation window from to 5 years by restricting the articles to those from 2014-15 changes the correlations little but increases the confidence interval widths (not shown).
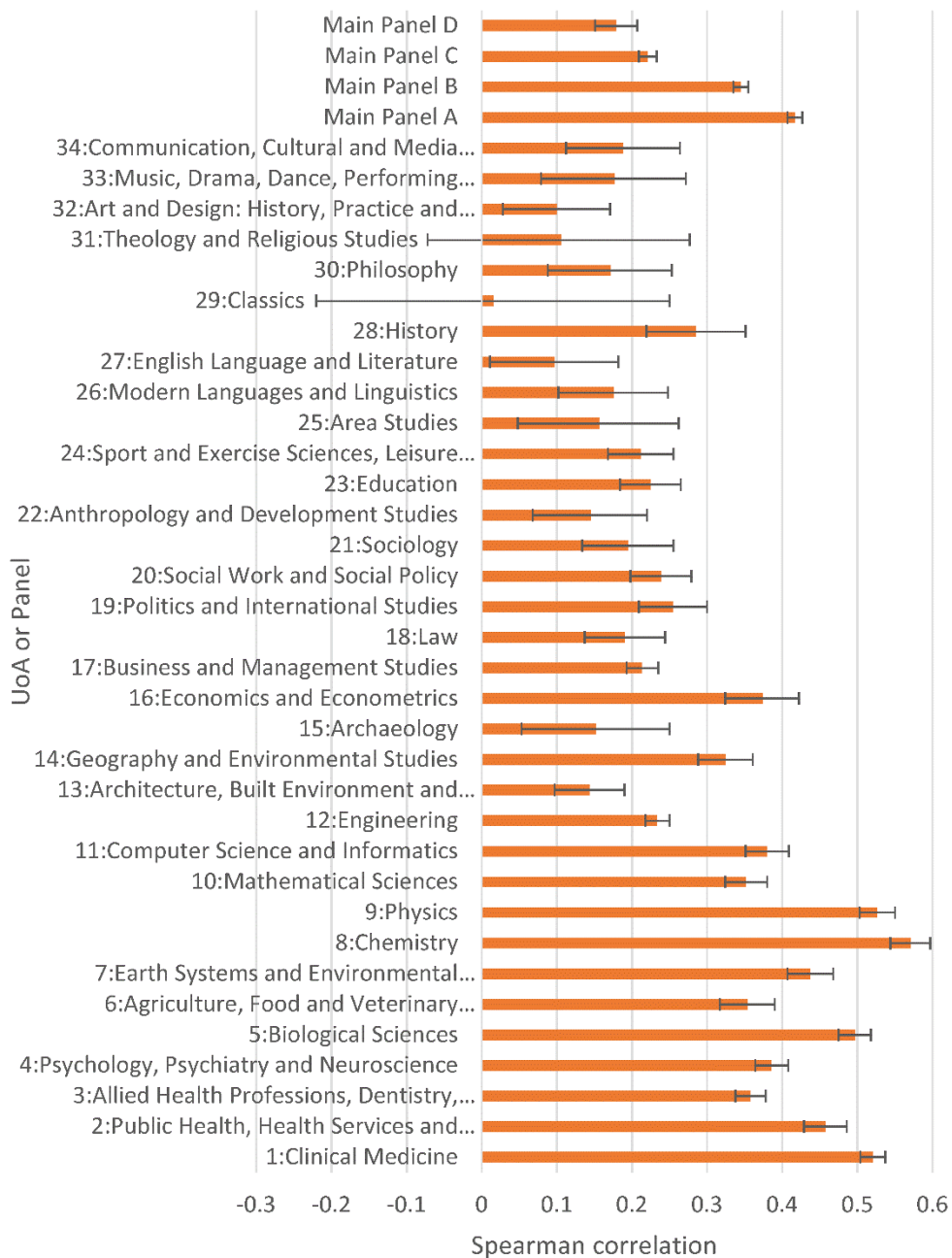
Figure 1. Spearman correlations between field and year normalised citation counts (NLCS) and REF scores for 2014-18 journal articles submitted to UK REF2021 by submitting Unit of Assessment or Main Panel. Error bars indicate 95% confidence intervals.
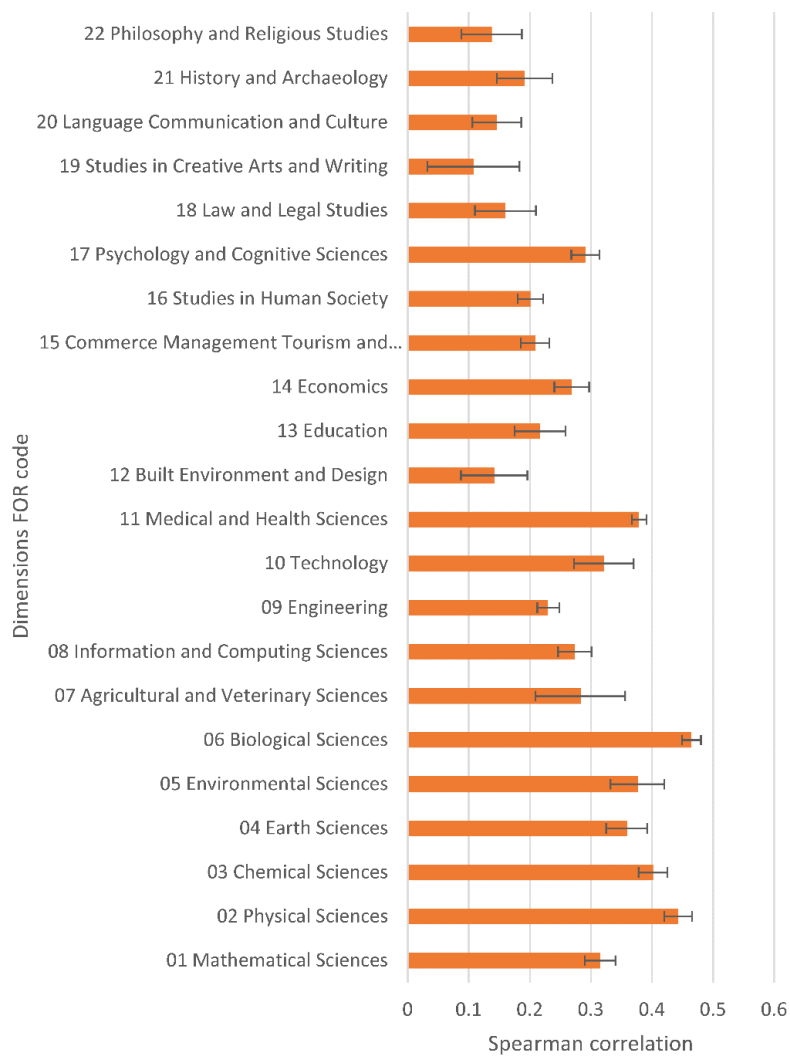
Figure 2. Spearman correlations between field and year normalised citation counts (NLCS) and REF scores for 2014-18 journal articles submitted to UK REF2021 by Dimensions FOR code (n=22). Error bars indicate 95% confidence intervals.
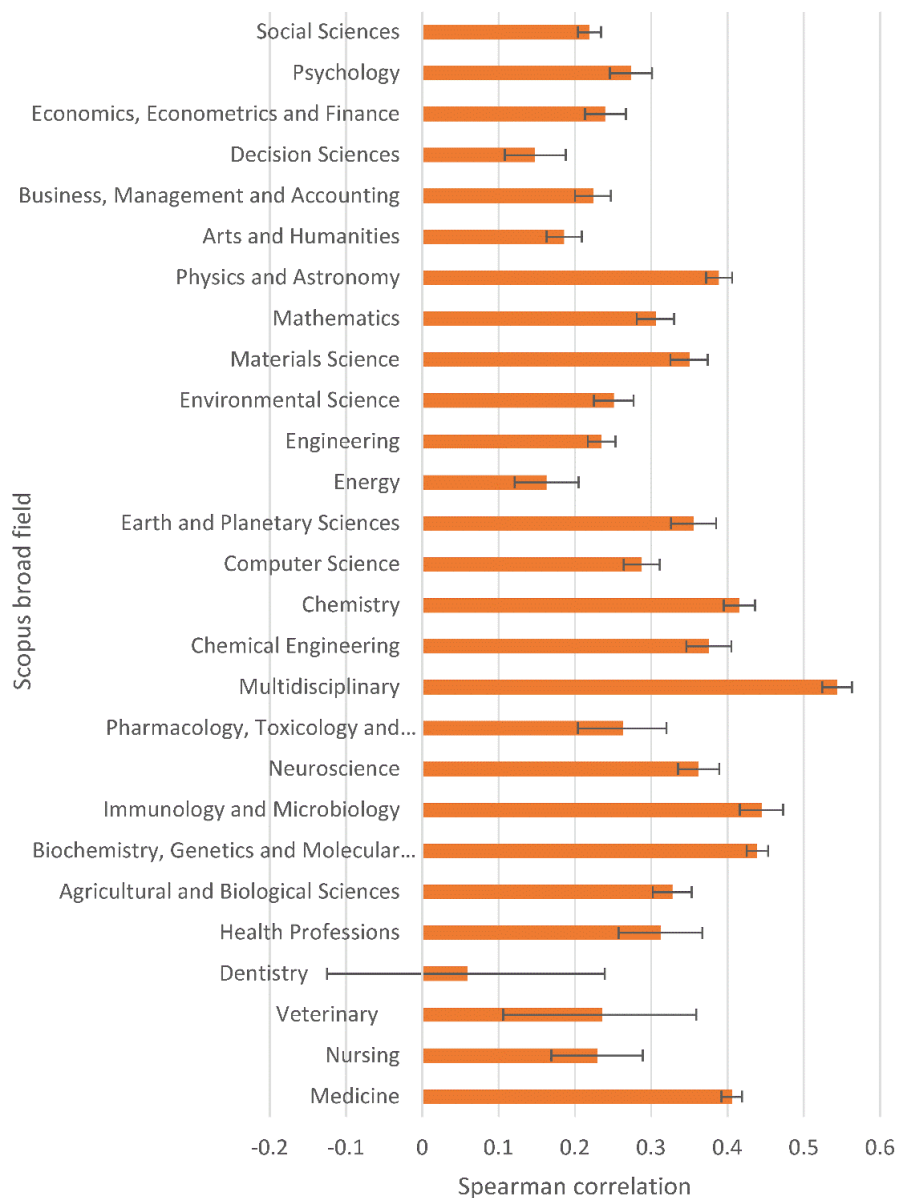
Figure 3. Spearman correlations between field and year normalised citation counts (NLCS) and REF scores for 2014-18 journal articles submitted to UK REF2021 by Scopus broad field (n=27). Error bars indicate 95% confidence intervals.

3.2 RQ3: Do high citations guarantee high quality in any field

In answer to RQ3, after bucketing articles together into groups of size at least 25, there were no UoAs in which the top group all had the highest REF2021 quality score (Figure 4). Thus, the simple answer to the research question is no: there is no citation threshold in any UoA that guarantees the highest quality score. Seven UoAs are close to this, however. Increasing the citation window to 3 years (articles from 2014-17, citations from 2021), four years (articles from 2014-16, citations from 2021) or five years (articles from 2014-15, citations from 2021) does not change the result: the most cited (NLCS) 25 articles in each UoA are never always rated 4*.
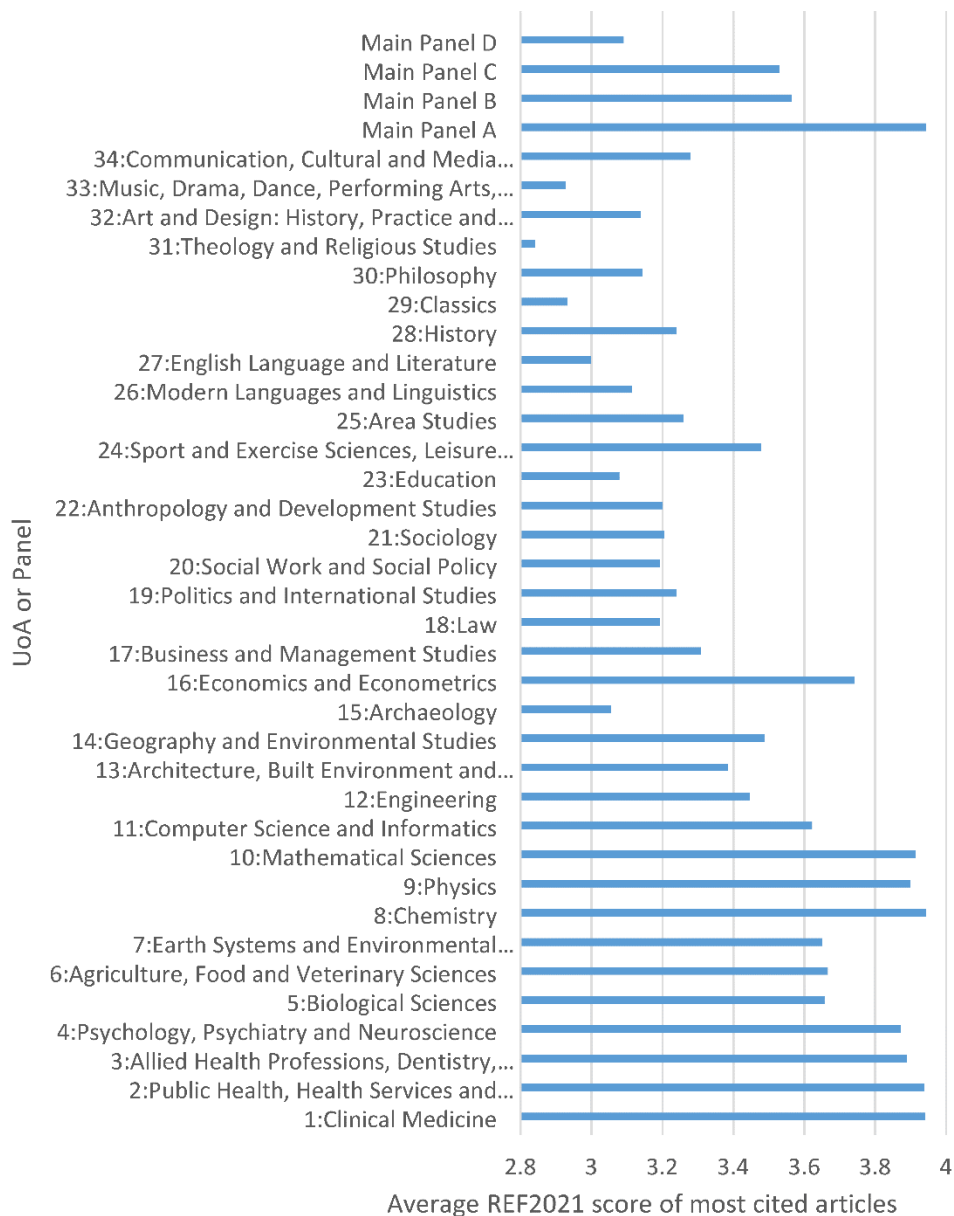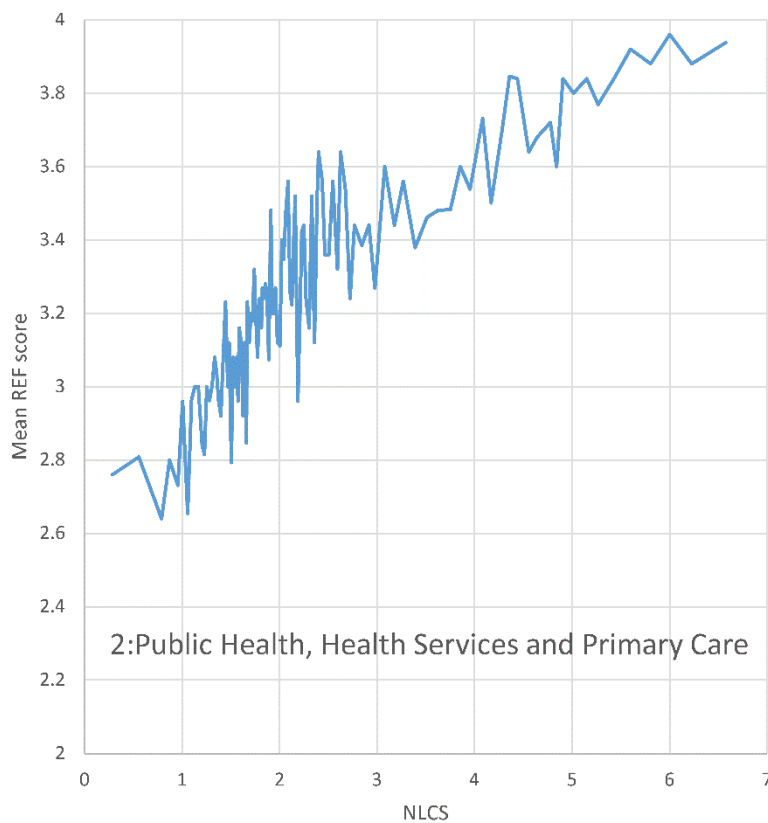
Figure 4. Mean REF scores for 2014-18 journal articles submitted to UK REF2021 for the 25 articles with the highest field and year normalised citation counts (NLCS) by submitting Unit of Assessment or Main Panel.

3.3 RQ4: Overall shape of the relationship between citations and research quality

A positive Spearman correlation can reflect many different underlying shapes, so it is informative to examine the underlying relationship between citations and research quality. The clearest way to do this is to plot average REF2021 scores against NLCS values, bucketing articles into groups with similar NLCS and taking the mean REF2021 score. This hides the variation between articles with similar NLCSs but shows the underlying trend. This is a problematic approach because the scores 1* to 4* are ordered but not scalar. Nevertheless, it is at least plausible to interpret 1* to 4* as a scale 1 to 4 and this assumption is routinely made for departmental Grade Point Averages (GPAs) constructed from REF scores. Given that this aspect of the calculation of GPAs does not seem to be challenged in the UK, it seems reasonable to make the same assumption here.

In all cases where there is a positive correlation above 0.1, the underlying shapes are close to straight lines, but some are more are consistent with approximate logarithmic curves: relatively rapid increase in average REF scores for NLCS increases at lower NLCS values and smaller increases in average REF scores for NLCS increases at higher NLCS values. The steepness of the increase and the range of average REF scores differs substantially between UoAs, however.

In fields with higher correlations (e.g., Figures 5, 6), the increase in average REF score for higher NLCS values is relatively steep, ending close to 4. Although there are variations within each NLCS range, in these fields, citation counts seem to be good indicators of quality and it would be surprising to find an excellent little cited article or a non-excellent highly cited article.

4

Figure 5. Mean REF scores for 2014-18 journal articles submitted to UK REF2021 in UoA 2 against field and year normalised citation counts (NLCS). Articles are bucketed into groups of at least 25 with similar NLCS.
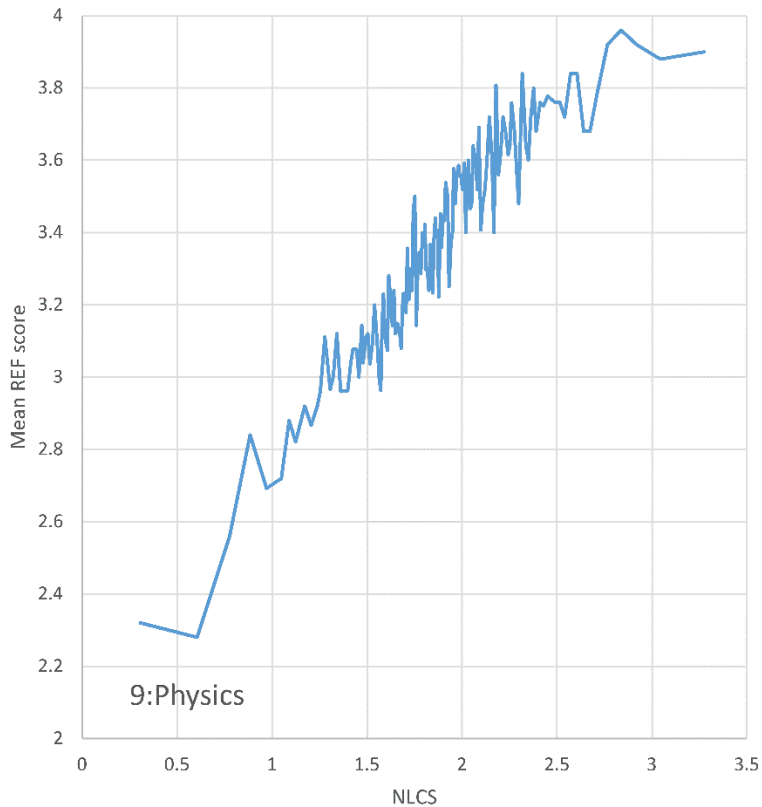
Figure 6. Mean REF scores for 2014-18 journal articles submitted to UK REF2021 in UoA 9 against field and year normalised citation counts (NLCS). Articles are bucketed into groups of at least 25 with similar NLCS.

In UoAs where the correlation between NLCS and REF scores is more moderate, the slope of the trend between NLCS and average REF2021 scores is less steep but still clear and does not get as close to the maximum (Figures 7, 8). In these fields, whilst there is a tendency for more cited articles to be higher quality, many articles break this trend.
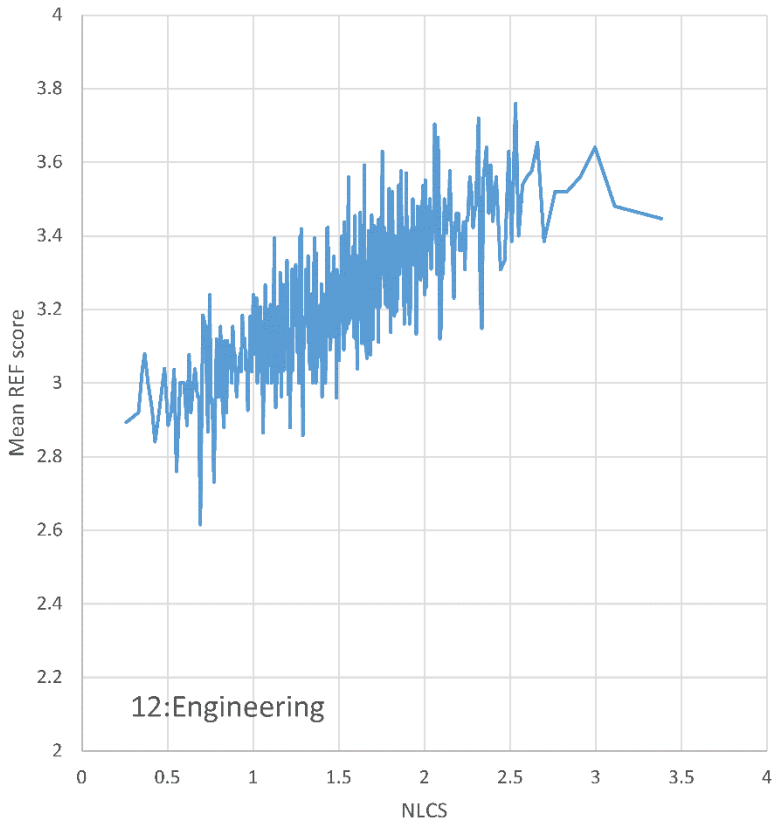
Figure 7. Mean REF scores for 2014-18 journal articles submitted to UK REF2021 in UoA 12 against field and year normalised citation counts (NLCS). Articles are bucketed into groups of at least 25 with similar NLCS.

Figure 8. Mean REF scores for 2014-18 journal articles submitted to UK REF2021 in UoA 17 against field and year normalised citation counts (NLCS). Articles are bucketed into groups of at least 25 with similar NLCS.
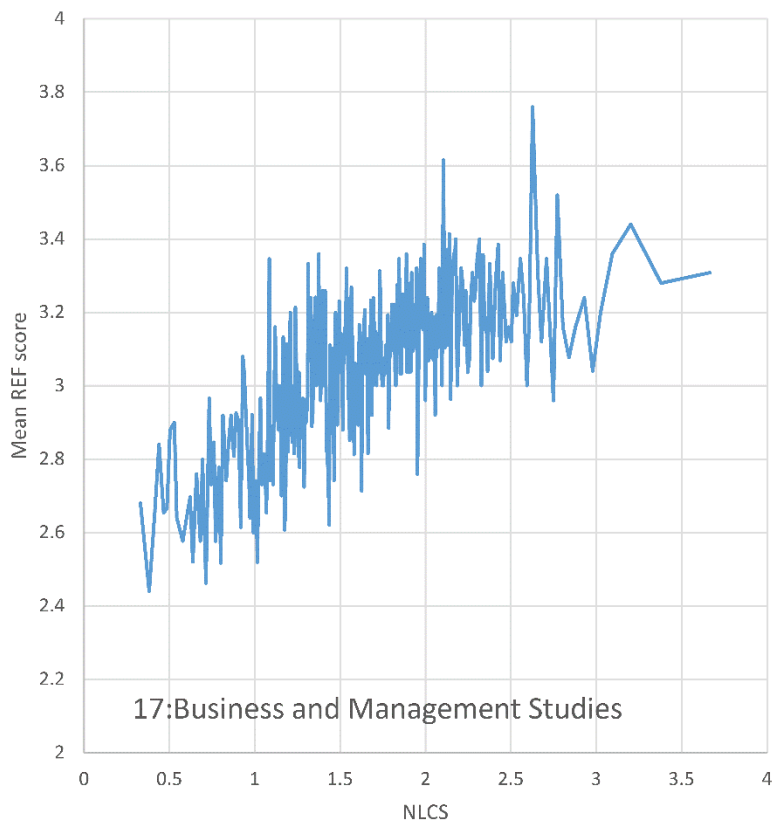
In UoAs where the correlation between REF2021 scores and NLCS is close to 0, this probably reflects a very shallow increasing tendency rather than a more complex relationship (e.g., not a U-shaped curve) (Figures 9, 10). A shallow curve like that in Figure 9 may reflect combinations of fields, some of which have no relationship between citations and quality (e.g., modern languages) and others that have some relationship (e.g., computational linguistics).
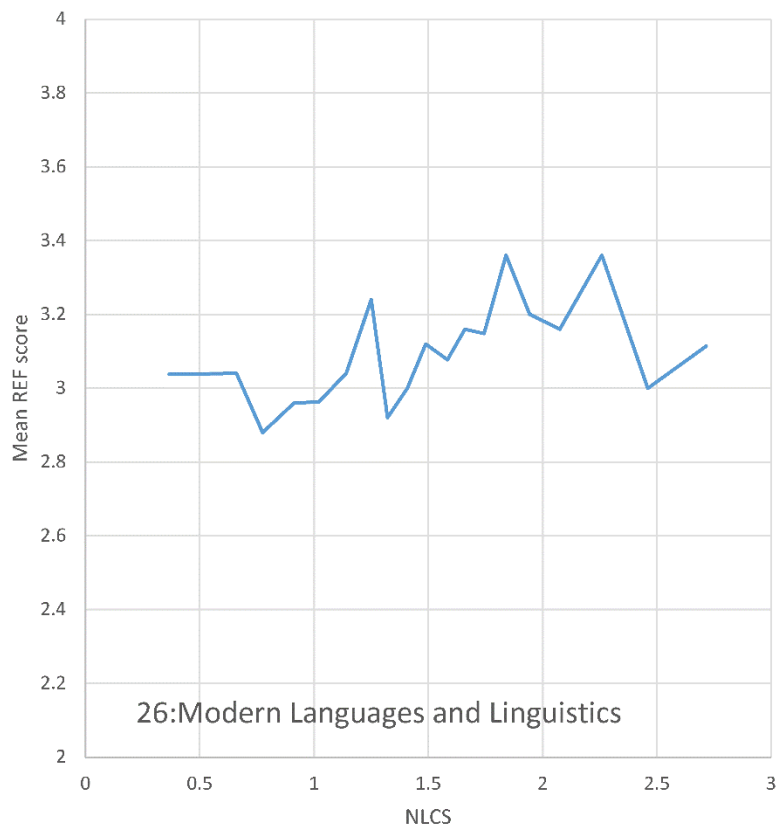

Figure 9. Mean REF scores for 2014-18 journal articles submitted to UK REF2021 in UoA 26 against field and year normalised citation counts (NLCS). Articles are bucketed into groups of at least 25 with similar NLCS.
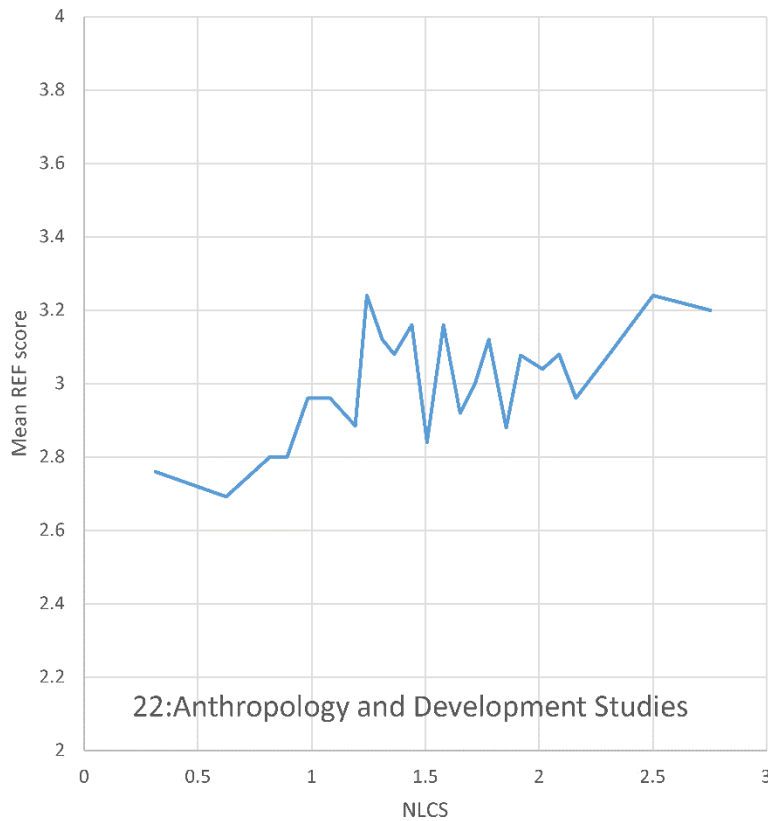
Figure 10. Mean REF scores for 2014-18 journal articles submitted to UK REF2021 in UoA 22 against field and year normalised citation counts (NLCS). Articles are bucketed into groups of at least 25 with similar NLCS.

5   Discussion

Although this is the largest study of its kind it has many limitations. First, all journal articles are from the UK and the relationship between citations and quality (and its different operationalisations) might be different in other countries, such as those that value research applications more highly than scientific contributions. Second, the articles are self-selected and represent the outputs considered by the authors to be their best work. The relationship might be different for lower quality research. Third, the field normalisation is limited by the primarily journal-based categorisation scheme of Scopus, which might generate anomalies through interdisciplinary journals. Fourth, there may well be narrow fields (and other output types) for which the relationship between citations and research quality is inverted or null. Finally, the bucket size used to investigate RQ3 and RQ4 is relatively arbitrary.

The results update and extend the largest prior document-level related investigation (HEFCE, 2015), by showing for the first time with extra statistical power and field classification systems, that a positive relationship between research quality and citations is relatively universal. It was already known that the strength of the relationship varied between fields at the institutional level (e.g., Franceschet & Costantini, 2011; HEFCE, 2015, Mahdi et al., 2008) and for articles (HEFCE, 2015), but not its universally positive nature. Although not all correlation confidence intervals excluded zero, the correlations were positive for all 34 UoAs, all 22 FOR codes and all 27 Scopus broad fields. Out of these, only three confidence intervals contained 0 and these were all for small sample sizes. Thus, whilst not fully proven, the results are consistent with a positive relationship occurring across all broad academic fields, and give strong evidence that the relationship is near universal. The statistical power of the large

numbers of articles in many fields supports this conclusion even for fields where the correlation is weak.

The finding that there is no reasonable citation threshold (field and year normalised) above which all articles are world leading research confirms that citation counts are never fully effective substitutes for human judgement, even in extreme cases. Whilst it is well known that articles occasionally become highly cited for negative reasons (e.g., the MMR/autism study: Godlee et al., 2011; the cold fusion article: Berlinguette et al., 2019), the results suggest that is it normal for articles in all fields to occasionally become extremely highly cited without having world leading quality. Moreover, in many fields (most UoAs) an extremely highly cited article is likely to be *not* world leading (e.g., averages below 3.5 in Figure 4). This does not undermine the use of percentiles in research evaluation, such as reporting the percentage of articles in the most cited 1% for a country (e.g., Rodriguez-Navarro, & Brito, 2022) but it cautions against fully equating highly cited with world leading research in any fields at the individual article level.

The close to linear relationship between the field and year normalised citation counts and research excellence is apparently the first finding of its kind. It should not be interpreted at face value for two reasons, however. First, REF scores are ordinal rather than forming a scale: it is not clear that the gap between, say, 1* and 2* is the same as the gap between 3* and 4*, or even that the concept of gap width in this context is meaningful. In the absence of evidence to the contrary, it is reasonable to at least hypothesise that the scores form a numerical scale. Nevertheless, the citation counts are log transformed as part of the NLCS calculation, so the x axis of Figure 5 to 10 is effectively log-transformed. If the x-axes were reverse log transformed, expanding the difference between the higher numbers, then the graph shapes would be close to logarithmic. Thus, it is reasonable to hypothesise that the underlying relationship between research quality and citation counts is logarithmic, with citation counts providing diminishing returns in terms of increased probability of higher quality at higher values. This would fit with the rich-get-richer phenomenon by which highly cited articles are believed to attract new citations partly because they are highly cited rather than for their intrinsic value (Merton, 1968).

6   Conclusions

The universal positive association between citation counts and research quality should provide reassurance for those that appropriately use citation-based indicators to support research quality evaluations. They also suggest, unexpectedly, that there are no broad fields of scholarship for which citations are completely irrelevant. Nevertheless, the wide variation between fields in the strength of the relationship confirms that citation-based indicators need greater levels of aggregation to yield useful information in some fields than others. For example, in fields with correlations above 0.5 at the article level, very strong aggregate correlations between average citations and average quality might be expected for small departments or small journals whereas the same aggregate correlations might only appear for very large departments or very large journals in other fields. Thus, the argument against inappropriate use of citations should not be that they are completely irrelevant in a field but that it is not reasonable to use them at a too low level of aggregation. Of course, if there are systematic bias in the citation data that field normalisation cannot eliminate, such as against qualitative research in a mixed methods field, then citation-based indictors would need to be used very cautiously in any context.

The fact that extremely high citation counts do not guarantee the highest research quality in any field and are not a high probability indicator of it in most (at least at the level of REF2021 UoAs) is a new finding. This should be remembered when journal articles are ranked by citations to identify the most influential articles in a field (Shadgan et al., 2010). For example, in June 2022 Google Scholar reported 302 articles containing the phrase "top cited articles" and 69 for "top cited papers" and many other bibliometric investigations include lists of top cited articles even if they do not focus on them. In this context, it should always be remembered that articles can become highly cited for reasons other than research excellence.

FUNDING INFORMATION

DATA AVAILABILITY

Extended information about the data can be found in an associated report (http://cybermetrics.wlv.ac.uk/TechnologyAssistedResearchAssessment.html). The raw data was deleted before submission to follow UKRI policy for REF2021.

7 References

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2011). National research assessment exercises: a comparison of peer review and bibliometrics rankings. Scientometrics, 89(3), 929-941.

Abramo, G., & D'Angelo, C. A. (2016). Refrain from adopting the combination of citation and journal metrics to grade publications, as used in the Italian national research assessment exercise (VQR 2011–2014). Scientometrics, 109(3), 2053-2065.

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.

Baccini, A., & De Nicolao, G. (2016). Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise. Scientometrics, 108(3), 1651-1671.

Berlinguette, C. P., Chiang, Y. M., Munday, J. N., Schenkel, T., Fork, D. K., Koningstein, R., & Trevithick, M. D. (2019). Revisiting the cold case of cold fusion. Nature, 570(7759), 45-51.

Borgman, C. L., & Furner, J. (2002). Scholarly communication and bibliometrics. Annual review of information science and technology, 36(1), 1-53.

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.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4), 507-521.

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

Godlee, F., Smith, J., & Marcovitch, H. (2011). Wakefield's article linking MMR vaccine and autism was fraudulent. *Bmj*, 342.

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/

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., & Rafols, I. (2015). Bibliometrics: the Leiden Manifesto for research metrics. Nature, 520(7548), 429-431.

Hicks, D. (2004). The four literatures of social science. Handbook of Quantitative Science and Technology Research, 473-496.

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.

Lyu, D., Ruan, X., Xie, J., & Cheng, Y. (2021). The classification of citing motivations: a meta-synthesis. Scientometrics, 126(4), 3243-3264.

MacRoberts, M. H., & MacRoberts, B. R. (2010). Problems of citation analysis: A study of uncited and seldom-cited influences. Journal of the American Society for Information Science and Technology, 61(1), 1-12.

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

McKiernan, E. C., Schimanski, L. A., Nieves, C. M., Matthias, L., Niles, M. T., & Alperin, J. P. (2019). Meta-research: Use of the journal impact factor in academic review, promotion, and tenure evaluations. Elife, 8, e47338.

Merton, R. K. (1968). The Matthew Effect in Science: The reward and communication systems of science are considered. Science, 159(3810), 56-63.

Merton, R. K. (1973). The sociology of science: Theoretical and empirical investigations. Chicago, IL: University of Chicago Press.

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.

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

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.

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). Berlin, Germany: Springer.

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

REF2021 (2020). Index of revisions to the 'Panel criteria and working methods' (2019/02). https://www.ref.ac.uk/media/1450/ref-2019_02-panel-criteria-and-working-methods.pdf

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.

Rodriguez-Navarro, A., & Brito, R. (2022). The extreme upper tail of Japan's citation distribution reveals its research success. arXiv preprint arXiv:2201.04031.

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

Shadgan, B., Roig, M., HajGhanbari, B., & Reid, W. D. (2010). Top-cited articles in rehabilitation. Archives of physical medicine and rehabilitation, 91(5), 806-815.

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

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

van Raan, A. F. (2004). Measuring science. In Handbook of quantitative science and technology research (pp. 19-50). Springer, Dordrecht.

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.

Vinkler, P. (1987). A quasi-quantitative citation model. *Scientometrics*, 12(1-2), 47-72.

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.

Whitley, R. (2000). The intellectual and social organization of the sciences. Oxford, UK: Oxford University Press.

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., (2015). The metric tide. Report of the independent review of the role of metrics in research assessment and management.