Terms in journal articles associating with high quality: Can qualitative research be world-leading?[1]

Mike Thelwall, Kayvan Kousha, Mahshid Abdoli, Emma Stuart, Meiko Makita, Paul Wilson, Jonathan Levitt.
Statistical Cybermetrics and Research Evaluation Group, University of Wolverhampton, UK.

Purpose: Scholars often aim to conduct high quality research and their success is judged primarily by peer reviewers. Research quality is difficult for either group to identify, however, and misunderstandings can reduce the efficiency of the scientific enterprise. In response, we use a novel term association strategy to seek quantitative evidence of aspects of research that associate with high or low quality.
Design/methodology/approach: We extracted the words and 2–5-word phrases most strongly associating with different quality scores in each of 34 Units of Assessment (UoAs) in the Research Excellence Framework (REF) 2021. We extracted the terms from 122,331 journal articles 2014-2020 with individual REF2021 quality scores.
Findings: The terms associating with high- or low-quality scores vary between fields but relate to writing styles, methods, and topics. We show that the first-person writing style strongly associates with higher quality research in many areas because it is the norm for a set of large prestigious journals. We found methods and topics that associate with both high- and low-quality scores. Worryingly, terms associating with educational and qualitative research attract lower quality scores in multiple areas. REF experts may rarely give high scores to qualitative or educational research because the authors tend to be less competent, because it is harder to make world leading research with these themes, or because they do not value them.
Originality: This is the first investigation of journal article terms associating with research quality.
Keywords: Research assessment; research quality; REF 2021; Research Excellence Framework; term frequency analysis; bibliometrics.

Introduction

Academic research in increasingly many countries is evaluated by post-publication peer review for quality assurance, formative feedback or to direct research funding (Engels & Guns, 2018; Woelert & McKenzie, 2018; Sivertsen, 2018; Jeon & Kim, 2018; Nielsen, 2017). The results can influence the reputations, funding, actions, and careers of the researchers involved. It is therefore important to investigate any systematic causes of high- or low-quality scores to find areas of good or bad practice and to check for bias. This is inherently difficult for academic peer review because even experts can strongly disagree on what is good research. Thus, whilst quality differences or biases can be tested for in researcher characteristics (e.g., gender, career status) or institutional status (e.g., prestige, size, geographic location), it is difficult to identify content-related patterns, such as research topics or methods that tend to attract high or low scores. The primary difficulty is that each research article is unique and not flagged with its core characteristics. Perhaps the most relevant general data for an article is its set of keywords, but not all articles have these, authors use

[1] Thelwall, M., Kousha, K., Abdoli, M., Stuart, E., Makita, M., Wilson, P. & Levitt, J. (in press). Terms in journal articles associating with high quality: Can qualitative research be world-leading? Journal of Documentation. https://doi.org/10.1108/JD-12-2022-0261

them differently, and controlled vocabularies are not universal. In response, we import and adapt a social science word frequency analysis method, word association thematic analysis (Thelwall, 2021), and apply it to the titles, abstracts and keywords of the articles evaluated to explore for article content characteristics that associate with high or low research quality.

Research quality is usually characterised as combining rigour, originality, and societal/academic significance (Langfeldt et al., 2020; REF2021, 2020). Whilst rigour is relatively objective, originality is subjective (Sánchez et al., 2019) and all three components depend on the expertise of the evaluator. For example, a statistical expert might be more critical of the quantitative component of the methods but might not notice the unusual degree of care that a qualitative researcher has taken with participant safety. More generally, expert reviewers can be the most critical (Gallo et al., 2016). Reviewers might also be biased by gender (Ceci & Williams, 2011; Fox & Paine, 2019), nationality (Harris et al., 2019; Primack et al., 2009; Thelwall et al., 2021), ethnicity (Woolston, 2021), and prestige (Tomkins et al., 2017). Cognitive cronyism, in the sense of judging results from known specialism better, is widely suspected but with little evidence (Lee et al., 2013; Wang & Sandström, 2015), and it is possible that cognitive cronies are more critical because they are more expert (e.g., Gallo et al., 2016). Indirect support for the hypothesis can be found from evidence that academics are more impressed by journals from their own specialties than others (Serenko & Bontis, 2018). Academics tend to give lower quality ratings to articles with conclusions that conflict with their beliefs, at least in psychology (Hergovich et al., 2010), perhaps because they are more suspicious of them. All these factors might explain the low degree of agreement between peer reviewers in many contexts (Fogelholm et al., 2012; Jackson et al., 2011; Kravitz et al., 2010; Pier et al., 2018; Rothwell et al., 2000; c.f., Pina et al., 2015), but do not suggest any content-related factors that might partly determine the quality of published research.

Some content-based factors are also known to relate to the citation impact or quality of an article. Findings might be judged to be more important if they are positive or statistically significant (Easterbrook, et al., 1991; van Lent et al., 2014). Individual research methods associate with differing levels of average citation impact, which may relate to their quality. In particular, articles reporting interviews, case studies, focus groups and ethnographies tend to be less cited in most fields (Thelwall & Nevill, 2021). Conversely, research mentioning questionnaires (Fairclough & Thelwall, 2022) or structural equation modelling (Thelwall & Wilson, 2016) tends to be more cited. There are field-based exceptions to these trends, however (no citation difference in library and information studies: Jamali, 2018; a qualitative advantage in international business research, although it classed survey articles as qualitative: López-Morales et al., 2022). Within most fields there are probably highly cited topics (e.g., Kim et al., 2011; Sanchez, 2020; Savin & van den Bergh, 2021), although it is not clear whether such topics would be generally agreed to include above average quality research. Interdisciplinary research may receive lower scores if the evaluators expect it to meet all the quality criteria of its constituent fields, so discussion between reviewers is helpful to understand the work from a holistic perspective (Huutoniemi, 2012; Oviedo-García, 2016). There are also legitimate types of methods bias in quality assessments based on hierarchies of evidence in some health-related fields (Katz et al., 2019; Murad, et al., 2016). For example, other factors being equal, a double-blind placebo-controlled randomised control trial is methodologically far more rigorous than a professional opinion (Vere & Gibson, 2021), although it could be considered less original.

No prior study has explored which contents of articles associate with research quality rather than citation impact. Terms in article titles, abstracts and/or keywords are often used

instead to map research topics (e.g., Ravikumar et al., 2015). Moreover, keywords (Kim et al., 2011) and manually identified themes (Sanchez, 2020) have been checked for associations with citation impact. The frequency of words or two- or three-word phrases in titles, abstracts, and keywords of articles within Scopus narrow categories has also been used to find terms occurring associating with hot topics showing them to be more cited in most fields (Thelwall & Sud, 2021). Repetition of keywords in abstracts also associates with citation counts for education journals (Sohrabi & Iraj, 2017) and the presence of popular management information system keywords can more effectively predict highly cited papers (n=746) than journal (e.g., Journal Impact Factor and SCImago Journal Rank) or author (author's h-index, publications, or citations) features (Hu et al., 2020).

In response to the scarcity of general evidence of the relationship between research quality and article contents, we applied a modified version of word association thematic analysis to detect words and themes associated with high- or low-quality research, as evaluated in REF2021. This is an exploratory method in the sense that we tested no hypotheses. Instead, the method itself generates the words and themes that are its output. The following general research questions drive our analysis.

- RQ1: Which types of words or phrases in articles and titles associate with higher quality research, if any?
- RQ2: Does the answer to RQ1 vary between fields?
- RQ3: Do the answers to RQ1 and RQ2 have wider implications for research evaluation?

Methods

The REF score data was supplied as part of The Responsible Use of Technology-assisted Research Assessment project, organised by the four UK higher education funding bodies. We used 148,977 journal articles submitted to REF2021 by publicly funded higher education institutions, except the University of Wolverhampton, which were redacted. Each article had identifying information, such as the journal, title and DOI, as well as its provisional score, as of March 2022. The results were published in May 2022 and the provisional results are very close to the final values, according to the REF team that supplied them. For confidentiality reasons, we deleted the data on May 8, 2022.

Each REF2021 journal article had a quality score allocated through single blind post-publication peer review by subject experts (for a detailed description of the similar process in the previous REF, see: Pidd & Broadbent, 2015). The REF procedure was as follows. Each article had been submitted to one of 34 field-based UoAs, each of which had a group of senior researchers (mostly full professors) to evaluate submitted outputs, with over 1000 evaluators for the REF overall. Articles were assigned by subpanel (UoA) chairs to the two evaluators judged to have the most relevant expertise. These evaluators independently scored outputs, resolved any disagreements and then their final scores were ratified by the entire subpanel. The judgements were single scores encompassing originality, significance, and rigour, with general guidelines (e.g., "Originality will be understood as the extent to which the output makes an important and innovative contribution to understanding and knowledge in the field."; "Significance will be understood as the extent to which the work has influenced, or has the capacity to influence, knowledge and scholarly thought, or the development and understanding of policy and/or practice"; and "Rigour will be understood as the extent to which the work demonstrates intellectual coherence and integrity, and adopts robust and appropriate concepts, analyses, sources, theories and/or methodologies") and some subject-specific criteria (e.g., "agenda-setting" and "the scale, challenge and logistical difficulty posed

by the research" were mentioned for some UoAs but not others) (REF2021, 2020). There were norm referencing exercises to ensure that scores were consistent across each UoA. There were also public written guidelines for the key procedures and quality criteria (REF2021, 2020). Overall, this was a careful exercise using guided expert judgment that was designed to give scores with a high degree of confidence and consistency. Nevertheless, the system is imperfect. The evaluators did not have to produce public written justifications for their scores (unlike journal article reviewers), may have had their own topic prejudices or other biases (e.g., Cotton et al., 2018; Lee et al., 2013; Haffar et al., 2019), and may not have devoted sufficient time to read and understand all outputs assigned to them. In addition, an unknown proportion of the articles submitted to the REF had no evaluators capable of understanding them and so their scores would have had to be guessed.

We removed 318 unclassified articles before the analysis. The REF records did not contain article titles abstracts and keywords, so we matched the articles to Scopus 2014-2020 for these. We searched REF outputs by DOI in a local copy of Scopus, generating 133,218 matches. We automatically searched the remaining articles by title and journal name (after converting to lower case and removing spaces), and manually checked the results to filter out false matches (typically articles with sort generic titles, such as, "Comment") to give an additional 997 results. We removed additional copies of articles that had been submitted by multiple institutions to the same UoA (or Main Panel for the panel-based analysis). For duplicate articles in the same UoA or Main Panel, we used the median score or one of the two medians at random when there was a tie. We analysed the articles extracted primarily by UoA to give the finest grained results, with UoAs grouped into four Main Panels and one complete set to identify more general trends. We removed articles scoring 0 since these or their authors may have been out of scope. We also removed articles with abstracts shorter than 500 characters, after cleaning, because these seemed to be a different type of output, such as a letter or comment, and therefore not comparable (Table 1).

Table 1. The number of journal articles submitted to the REF and matching Scopus 2014-2020, after removing duplicates and removing articles with cleaned abstracts shorter than 500 characters.

UoA or Panel	Articles
1	9905
2	3889
3	9675
4	8172
5	6376
6	3147
7	3724
8	3274
9	4499
10	5111
11	4645
12	16333
13	2582
14	3439
15	545
16	1762
17	11851
18	1864
19	2502
20	3294
21	1498
22	977
23	3336
24	2812
25	524
26	962
27	768
28	1082
29	111
30	806
31	185
32	1117
33	544
34	1020
Main Panel A	37282
Main Panel B	36584
Main Panel C	35631
Main Panel D	7071
UoA total	**122331**
Panel total	**116568**

We used chi square tests to identify words that occurred disproportionately often in articles with different quality levels. First, we cleaned article abstracts for journal standard texts, such

as copyright statements, structured headings, and open access statements. This cleaning was automatic using a large set of heuristics initially created for a previous study (Thelwall & Sud, 2021) and updated for our programme of work on the REF outputs. After the data cleaning, we extracted words and phrases of up to five words not spanning sentence boundaries from each article. A limit of five words in a phrase was set to capture relevant short phrases without overwhelming the results with longer, over-specific terms. We merged the lowest two scores (1* and 2*) into a single group to increase statistical power since the 1* group was very small. In word frequency analyses it is common to remove stopwords (very common short words like "the" and "a") to increase the statistical power of the tests. We did not do this because some common words seemed to be useful to retain (e.g., "we", "our"), as shown by the results. In retrospect, removing stopwords would have allowed the results to focus more on research topics and methods and may have generated more useful results, however.

In each UoA and Main Panel, we calculated a chi square value for every word and phrase extracted to assess whether it occurred disproportionately often in one of the three quality groups (1* or 2*, 3*, 4*). This value reflects whether the three quality classes have different proportions of articles containing each term. For example, if 1% of 1* or 2* articles contained "funded by", 2% of 3* articles contained "funded by" and 5% of 4* articles contained "funded by" then these differences would translate into a large chi square value, and if the percentages were identical (e.g., all 2%) then the chi square value would be 0.

For each UoA and main panel, we examined the fifty terms with the highest chi squared values. Since spurious statistical positives can occur with multiple significance tests, a Bonferroni correction (Ranstam, 2016) was applied to identify a minimum chi square value (i.e., a corrected α=0.05 significance level) in each UoA or Main Panel for a term to be statistically significant. We ignored terms with a chi square below this value, even if they were in the top 50 for the UoA or Main Panel. In cases where this resulted in no statistically significant terms for a UoA, we added terms with the highest chi square values for illustrative purposes and flagged them as such.

The words and phrases were grouped into three themes according to their main apparent purpose, following the word association thematic analysis method (Thelwall, 2021). This purpose was identified by reading the contexts of the words and phrases in matching article titles, keywords, or abstracts, iteratively grouping the contexts into themes, and re-checking the initial contexts until a final set of themes emerged. The three themes we found were style, methods, and topic. Although the main three themes seemed clear, some words or phrases matched multiple themes, so the final placements were subjective. For example, "we demonstrate" was classified as stylistic but the marginally more specific "we prove that" and "we identify" were classified as methods even though both have stylistic elements (first person singular). The first author performed the classifications. The use of only one coder was an unfortunate necessity due to the time limited access to the data.

Results

Main Panel A: Words and phrases

For Main Panel A, which mostly focuses on life sciences, health and medicine, there are many stylistic terms that associate with prominent journals (Table 2). In particular, "here we show that" is a common phrase in enough prestigious journals in this area to statistically associate with high quality research. This phrase occurs in abstracts, usually in the middle after introducing the context of the study, but sometimes at the start. More generally, first person

plural singular (we, our) in the present tense is a common style in several prestigious journals in this area (e.g., *Blood, Cell*), whereas the third person and past tense more associate with other journals (e.g., "this study was"). For example, whilst 60.8% of REF journal article extracts contained "we", it was in 99% of abstracts in the journal *Nature,* 96% of *Science,* 94% of *The Lancet,* and 91% of *Cell*. This suggests that it is an actual or de facto style requirement for these journals.

Funding associated with higher quality in most UoAs in this Main Panel, typically through a declaration at the end of an abstract, such as, "Funding: This project was funded by the NIHR." This seems to be a journal style issue because funding was mainly mentioned in *The Lancet* and its family of journals (e.g., *The Lancet Public Health*). It is unlikely to be a funding issue since most studies in this area were presumably externally funded.

Methods terms partly reflect standard hierarchies of evidence or at least indicators of higher quality studies (e.g., double-blind, masked, "randomly-assigned patients"). In some UoAs, qualitative methods associate with lower quality (e.g., interviews, themes, thematic, qualitative) although the term "measured" also associated with lower quality in one. The suggestion that qualitative research tended to score lower could be due to bias in favour of quantitative research or adherence to evidence hierarchies that do not include qualitative studies. Based on reading abstracts in UoA 3 containing "themes" or "thematic", since qualitative studies are based on limited samples, they provide insights and may trigger suggested actions (e.g., "Resources are needed that are tailored to men, framed around fatherhood") but it might be more difficult to argue that the findings are world leading because the conclusions seem unlikely to have direct societal impact or to be definitive. This is possibly a generic issue with evaluating qualitative research.

In some UoAs there were terms indicating topics that tended to be higher or lower quality, although there is also an overlap with methods. For example, mice and mouse models are used in many research methods. The topics may associate with the main areas of strong or weak research groups submitted to each UoA rather than being intrinsically more important.

Table 2. Examples of words and phrases with the strongest associations (chi-square test) with REF scores by UoA for Main Panel A. Bold terms associate with lower REF scores; other terms associate with higher REF scores. Words and shorter phrases within longer phrases are omitted in favour of the longest relevant phrase. Words or phrases matching multiple themes are listed only under their apparent main theme.

UoA	Style	Methods	Topic
1: Clinical Medicine	We, here, **were**	Funding, "randomly assigned patients", "the primary outcome", double-blind, interpretation, masked, "trial is registered with", "in the placebo group", "group and", "to receive", "primary outcome", intention-to-treat, "adverse events"	
2: Public Health, Health Services and Primary Care	"This is an"	Funding, "randomly assigned", "the primary outcome", randomisation, interpretation, trial, "trial is registered", "adverse events", "group and", **interviews, participation**	
3: Allied Health Professions, Dentistry, Nursing and Pharmacy	"Here we", "we show that", **was, were**	Funding, CI, "trial is registered", "adverse events", randomised, randomisation, "randomly assigned", "the primary outcome was", intention-to-treat, stratified, "adverse events", **themes**, **thematic**	
4: Psychology, Psychiatry and Neuroscience	"Here we show that", **were, was, "the aim", "the current study", "the aim"**	Funding, "randomly assigned", "trial is registered", "the primary outcome was", vivo, **online, web, "measures of", completed, discussed, research, qualitative, interviews**	Neurons, neuronal, gene, human, mouse, cell, protein, brain, synaptic, disease, **participants**
5: Biological Sciences	"Here we show that", **be, were, "this study was", "did not", no, investigated, conducted, some, "of this study", "used to", may, had, been, or**	**Web, "the effects of", "a significant", compared, assessed, "to assess", mean, measured**	
6: Agriculture, Food and Veterinary Sciences	"Here we show that", "we report", our, **were, "there was", significantly, "used to", on, no, had, "this study", "this paper", be,**	"We identify", replication, **collected, "evaluation of", "the effect of"**	"Evolution of", signaling, evolutionary, genes, cells, "Amino acid", mice, genome, genomic, "in arabidopsis", mechanism, **horses, "in dogs", "dogs with", farm**
Panel A	"Here we show that", **were, was, "this study"**	Funding, "we randomly assigned", "is registered with", "the primary outcome was", interpretation, "adverse events", "to receive"	Cells

Main Panel B: words and phrases

For Main Panel B (Table 3), there were similar journal style terms to Main Panel A. For methods, the results suggest that experimental work and proof tended to be rated higher quality and that qualitative research (again) tended to attract lower scores. Case studies are also mentioned for Main Panel B overall. This term could refer to the case study method or an investigation of a single example of something using other methods. The topics for Main Panel B UoAs could reflect both research group specialisms and important societal topics (e.g., warming, climate, ocean). Research mentioning students or higher education tended to be lower quality. Although students are sometimes used as a convenient population to study in non-educational research (e.g., in psychology) they were primarily investigated in the context of education in the articles examined for the table.

Table 3. Examples of words and phrases with the highest chi-square values by UoA for Main Panel B. Bold terms associate with lower scores; other terms associate with higher scores.

UoA	Style	Methods	Topic
7: Earth Systems and Environmental Sciences	"Here we show that", "we find that", "here we present", "**was also", were, "in this study", showed**	**Analysis, significant, investigated, method, behaviour, "compared to"**	"The global", warming, earth, climate, ocean, "million years", "years ago", atmospheric, ice, circulation, forcing, **UK Raw, wort**
8: Chemistry	"Here we", "we show that", **"were performed", was, showed, "were found to", "an investigation into"**	**Reduce, tests, assessment, evaluated, "the formulations" "and in vitro"**	
9: Physics	"Here we report", "so far",	"Has been developed"	
10: Mathematical Sciences	We	"We prove", **visualisations, recruited, "is in the use of", "trial registration", "and simplify the"**	**"Only a subset", "codes over rings", epidemiology, "the epidemic", aged, "from group rings"**
11: Computer Science and Informatics	"We show that", "we demonstrate that", "we introduce", our, **"this study", "the results", "this research", "this paper", project, "the results", presented**	Experiments, approximate, complexity, "we prove that", bounds, probabilistic, **review, interviews, development**	Problem, imaging, general, graph, first, polynomial, "class of", **technology, future**
12: Engineering	"Here we report/present/ demonstrate", "here we show that", **were, was, had, study, "results showed", "this paper", "the results", investigated, "carried out"**	"In vivo", **qualitative, interviews, analysis**	Imaging, photonic, quantum, optical, spatial
Panel B	"Here we show that", "here we demonstrate", "here we present", "we report", our, **was**, **were, "the purpose of this paper", "this study", "this research", used, "the results", showed, investigated, "the proposed", different**	"We prove", **analysis, "case study", compared, performance**	Warming, global, "years ago", "million years", quantum, **management, students, "higher education".**

Main Panel C: words and phrases

For Main Panel C, there were again style terms with a tendency for the first-person present tense to be higher quality and third person past tense to be lower quality (Table 4). This was again primarily journal-based. For example, in UoA 13 only 9 out of 100 *Energy and Buildings*

article abstracts contained “we” but it was in 39 out of 40 *Nature* family journal article abstracts. Qualitative methods were again given lower quality scores overall and for one UoA. Important global issues again scored well and, more clearly than before, education-related articles tended to attract lower scores (although not in UoA 23).

Table 4. Examples of words and phrases with the highest chi-square values by UoA for Main Panel C. Bold terms associate with lower scores; other terms associate with higher scores.

UoA	Style	Methods	Topic
13: Architecture, Built Environment and Planning	**"The purpose of this paper"**		
14: Geography and Environmental Studies	"Here we show", "we show that", "our results", "we find that", **"of this study", "the study", "the results", were, "there was a", used, showed**	"Per cent"	"In global", "a global", climate, oceanic, tropical, "earth system", **UK**
15: Archaeology	"Here we", "**this article"**,		Asia
16: Economics and Econometrics	We, "we develop", indicate, "**purpose of this paper is", "the authors", "based on an" "this study", "the findings", "willing to"**		**Annual, "for future", region**
17: Business and Management Studies	"We show", "we develop", when, "we find that", **"purpose of this paper is", "of this paper is to", "have been", was", "the findings", there**	"consistent with", **review, analysis**	Behaviour, **"the UK", "the period", policy, sector, crisis**
18: Law*	"We argue that"		**Students**
19: Politics and International Studies	"We find that", our, "we show that", **also, "this article"**	Data, effects, results, experiment	Electoral
20: Social Work and Social Policy	"We find that", "we use", our, show, **"the purpose of", researcher, research**	"Longitudinal study", long-term, panel, CI, "data for", cohort, effects, evidence, results, estimate, per, baseline, rates, modelling, **themes**	Household, birth, family, incentives, "the English", at age, **students, teaching, learning**
21: Sociology*	We	Models, "longitudinal study"	
22: Anthropology and Development Studies*	"We present"	Consistent	Modalities
23: Education	"We find"	Longitudinal, multilevel, measures, "**perceptions of", experiences**	Attainment, **staff, online**
24: Sport and Exercise Sciences, Leisure and Tourism	"Here we", "we show", **"this study"**	"Muscle biopsies were", "muscle protein synthesis", **"in** vivo", **"total distance", completed**	"Human skeletal muscle", "of muscle", "muscle mass", humans, expression, motor, atrophy, "**countermovement jump", "soccer players", sport**
Panel C	"we show that", "we find that", "here we", our, **"the purpose of this paper", "aims to", "the research", "this study", challenges**	"Consistent with", effects, **interviews, themes**	"Skeletal muscle", **"higher education", university, students, teaching, teachers. staff, experiences, issues, development**

* Listed terms are not statistically significant after a Bonferroni correction.

Main Panel D: words and phrases

No Main Panel D UoA had any statistically significant terms, but students are again mentioned overall as a lower quality topic (Table 5).

Table 5. Examples of words and phrases with the highest chi-square values by UoA for Main Panel D. Bold terms associate with lower scores; other terms associate with higher scores.

UoA	Style	Methods	Topic
25: Area Studies*			"The global", governments
26: Modern Languages and Linguistics*		Results	Semantics, **narrative**
27: English Language and Literature*	Whose		Phonetic, **"of lexical"**
28: History*	**Will**		"The old"
29: Classics*		**Concept, "reference to"**	**Wider**
30: Philosophy*	"It is"	Statistical, **"the historical"**	
31: Theology and Religious Studies*	"In some"	**Associated**	"The phenomenon"
32: Art and Design: History, Practice and Theory*	"This essay"	Examination	"And artistic"
33: Music, Drama, Dance, Performing Arts, Film and Screen Studies*	**My**	**Rated**	Musical, music, **performance**
34: Communication, Cultural and Media Studies, Library and Information Management*	Most		Senior, move, **"the local"**
Main Panel D			Syntactic, variation, **students, narrative**

* Listed terms are not statistically significant after a Bonferroni correction.

Discussion

The limitations of the results include that the journal articles all have at least one UK author, and are self-selected, with a cap of 5 per researcher. This may help authors in quantitative subjects that produce more work and can cherry pick their best outputs. Conversely, it may also help researchers that produce less work because each submitted output represents a larger share of their efforts. The REF system may also have pushed people that are primarily educators into conducting education-related research to participate, where they would compete with people that devote more time to research. If true, then the education research may have more value per quality point since it would have had less input. Moreover, the REF2021 rules may have affected the results as may any discussions within UoAs or main panels about the relative merits of different types of research. The results also rely on what is written in titles, abstracts and keywords, which may not translate directly to the topics of articles. For example, perhaps researchers that are more expert with qualitative research use more specific terms than "theme" or "qualitative", giving a second order quality effect for the remaining articles using these terms. The differences found may also be influenced by the output type. For example, qualitative research may work better in the longer format of monographs and book chapters, where it may have attracted higher scores.

The stylistic results are relatively uninteresting in the sense that they point to journal-based norms and differences in the average quality of journals are well known. It is therefore unsurprising that journal style norms translate into quality-associated stylistic terms. It is also possible that higher quality articles in other journals have similar stylistic features, either because the authors are experienced in submitting to prestigious journals, associate the style

with high quality research, or publish their best articles rejected from a prestigious journal to another outlet.

Some of the methods results for Main Panel A align with known hierarchies of evidence and good practice in medical fields by mentioning placebos, randomisation, double-blind, and trial registration (Vere & Gibson, 2021). Many of the other methods are quite specific and may relate to additional care taken with experiments or journal style guidelines about what to mention. The two main general results – the lower scores given to education-related research and qualitative research are different however, and do not seem to have been previously noted in studies of peer review bias. Nevertheless, there have been previous claims of general bias against qualitative research (Bansal & Corley, 2011) and those relying on hierarchies of evidence might regard it as being a low-level type (Vere & Gibson, 2021). Moreover, quantitative researchers used to larger sample sizes are known to sometimes devalue qualitative research for having few participants (Baillie & Douglas, 2014) or for lacking generalisability (Smith, 2018). Supporting this, interviews and case studies tend to be less cited (Thelwall & Nevill, 2021). No previous study seems to have remarked that educational research tends to get lower quality scores in research evaluation contexts, however, although concerns have been raised that the distinction between educational research and teaching and learning scholarship is blurred (Cotton et al., 2018), which may have resulted in some sub-standard REF submissions, or suspicion on the part of assessors. In addition, the capability of REF assessors for educational research has been questioned (Cotton et al., 2018).

Conclusion

The results show that there are stylistic methods and topic associations with different research quality scores for journal articles in most UoAs, especially those with large numbers of articles. Despite the focus on quality, no term directly mentioned an aspect of quality, such as through a claim to be novel, rigorous or impactful. The results suggests that there are common methodological associations with high scores, presumably because there are recognised hierarchies of method rigour. Since the style findings are journal-related and the individual topics could be due to individual research groups, the methods differences are the clearest general finding. Thus, a take-away message for researchers is to ensure that the most rigorous method is selected for each study.

The most worrying findings are the lower scores given in some UoAs to educational and qualitative research. As argued above, the former may be a systemic effect of the evaluation system; the latter is more concerning, given the need for methods pluralism in a healthy research system. The approach used here does not show whether qualitative research tended to be lower quality, whether the REF assessors tended be harsher on it, or whether the result is an artefact of the methods used (e.g., if weaker researchers were more likely to describe their approach as “qualitative”). Nevertheless, the suggestion that a major research approach may tend to attract lower scores is worrying. This is particularly important for the REF where, at the time of writing, only 4* research was fully funded, with 3* receiving 25% funding and the remainder nothing. At the moment, interdisciplinary research is given special consideration within the REF rules. Special consideration may also be needed for qualitative research to ensure that it is not undervalued in academia.

Acknowledgement

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.

References

Baillie, C., & Douglas, E. P. (2014), "Confusions and conventions: Qualitative research in engineering education", *Journal of Engineering Education*, Vol. 103 No. 1, pp. 1-7.

Bansal, P., & Corley, K. (2011), "The coming of age for qualitative research: Embracing the diversity of qualitative methods", *Academy of Management Journal*, Vol. 54 No. 2, pp. 233-237.

Ceci, S. J., & Williams, W. M. (2011), "Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences*, Vol. 108 No. 8, pp. 3157-3162.

Cotton, D. R., Miller, W., & Kneale, P. (2018), "The Cinderella of academia: Is higher education pedagogic research undervalued in UK research assessment?", *Studies in Higher Education*, Vol. 43 No. 9, pp. 1625-1636.

Doğan-Uçar, A., & Akbasb, E. (2022), "A corpus-driven cross-disciplinary study of inclusive and exclusive we in research article abstracts", *LEARN Journal: Language Education and Acquisition Research Network*, Vol. 15 No. 1, pp. 180-204.

Easterbrook, P. J., Gopalan, R., Berlin, J. A., & Matthews, D. R. (1991), "Publication bias in clinical research", *The Lancet*, Vol. 337 No. 8746, pp. 867-872.

Engels, T. C., & Guns, R. (2018), "The Flemish performance-based research funding system: A unique variant of the Norwegian model", *Journal of Data and Information Science*, Vol. 3 No. 4, pp. 45-60.

Fairclough, R., & Thelwall, M. (2022), "Questionnaires mentioned in academic research 1996–2019: Rapid increase but declining citation impact", *Learned Publishing*, Vol. 35 No. 2, pp. 241-252.

Fogelholm, M., Leppinen, S., Auvinen, A., Raitanen, J., Nuutinen, A., & Väänänen, K. (2012), "Panel discussion does not improve reliability of peer review for medical research grant proposals", *Journal of Clinical Epidemiology*, Vol. 65 No. 1, pp. 47-52.

Fox, C. W., & Paine, C. T. (2019), "Gender differences in peer review outcomes and manuscript impact at six journals of ecology and evolution", *Ecology and Evolution*, Vol. 9 No. 6, pp. 3599-3619.

Gallo, S. A., Sullivan, J. H., & Glisson, S. R. (2016), "The influence of peer reviewer expertise on the evaluation of research funding applications", *PloS One*, Vol. 11 No. 10, pp. e0165147.

Haffar, S., Bazerbachi, F., & Murad, M. H. (2019), "Peer review bias: a critical review", Mayo Clinic Proceedings, Vol. 94, No. 4, pp. 670-676.

Harris, M., Marti, J., Watt, H., Bhatti, Y., Macinko, J., & Darzi, A. W. (2017), "Explicit bias toward high-income-country research: a randomized, blinded, crossover experiment of English clinicians", *Health Affairs*, Vol. 36 No. 11, pp. 1997-2004.

Hergovich, A., Schott, R., & Burger, C. (2010), “Biased evaluation of abstracts depending on topic and conclusion: Further evidence of a confirmation bias within scientific psychology”, *Current Psychology*, Vol. 29 No. 3, pp. 188-209.

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*, Vol. 14 No. 1, pp. 101004.

Huutoniemi, K. (2012), “Communicating and compromising on disciplinary expertise in the peer review of research proposals”, *Social Studies of Science*, Vol. 42 No. 6, pp. 897-921.

Jackson, J. L., Srinivasan, M., Rea, J., Fletcher, K. E., & Kravitz, R. L. (2011), “The validity of peer review in a general medicine journal”, *PloS One*, Vol. 6 No. 7, pp. e22475.

Jamali, H. R. (2018), “Does research using qualitative methods (grounded theory, ethnography, and phenomenology) have more impact?”, *Library & Information Science Research*, Vol. 40 No. 3-4, pp. 201-207.

Jeon, J., & Kim, S. Y. (2018), “Is the gap widening among universities? On research output inequality and its measurement in the Korean higher education system”, *Quality & Quantity*, Vol. 52 No. 2, pp. 589-606.

Katz, D. L., Karlsen, M. C., Chung, M., Shams-White, M. M., Green, L. W., Fielding, J., ... & Willett, W. (2019), “Hierarchies of evidence applied to lifestyle Medicine (HEALM): introduction of a strength-of-evidence approach based on a methodological systematic review”, *BMC Medical Research Methodology*, Vol. 19 No. 1, pp. 1-16.

Kim, H. E., Jiang, X., Kim, J., & Ohno-Machado, L. (2011), “Trends in biomedical informatics: most cited topics from recent years”, *Journal of the American Medical Informatics Association*, Vol. 18 Supplement_1, pp. i166-i170.

Kravitz, R. L., Franks, P., Feldman, M. D., Gerrity, M., Byrne, C., & Tierney, W. M. (2010), “Editorial peer reviewers' recommendations at a general medical journal: are they reliable and do editors care?”, *PloS One*, Vol. 5 No. 4, pp. e10072.

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*, Vol. 58 No. 1, pp. 115-137.

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013), “Bias in peer review”, *Journal of the American Society for Information Science and Technology*, Vol. 64 No. 1, pp. 2-17.

López-Morales, J. S., Salazar-Núñez, H. F., & Zarrabal-Gutiérrez, C. G. (2022), “The impact of qualitative methods on article citation: an international business research perspective”, *Scientometrics*, Vol. 127 No. 3, pp. 3225-3236.

Murad, M. H., Asi, N., Alsawas, M., & Alahdab, F. (2016), “New evidence pyramid”, *BMJ Evidence-Based Medicine*, Vol. 21 No. 4, pp. 125-127.

Nielsen, M. W. (2017), “Gender consequences of a national performance-based funding model: new pieces in an old puzzle”, *Studies in Higher Education*, Vol. 42 No. 6, pp. 1033-1055.

Oviedo-García, M. Á. (2016), “Tourism research quality: Reviewing and assessing interdisciplinarity”, *Tourism Management*, Vol. 52, pp. 586-592.

Pidd, M., & Broadbent, J. (2015), “Business and Management Studies in the 2014 Research Excellence Framework”, *British Journal of Management*, Vol. 26 No. 4, pp. 569-581.

Pier, E. L., Brauer, M., Filut, A., Kaatz, A., Raclaw, J., Nathan, M. J., & Carnes, M. (2018), “Low agreement among reviewers evaluating the same NIH grant applications”, *Proceedings of the National Academy of Sciences*, Vol. 115 No. 12, pp. 2952-2957.

Pina, D. G., Hren, D., & Marušić, A. (2015), "Peer review evaluation process of Marie Curie actions under EU's seventh framework programme for research", *PLoS One*, Vol. 10 No. 6, pp. e0130753.

Primack, R. B., Ellwood, E., Miller-Rushing, A. J., Marrs, R., & Mulligan, A. (2009), "Do gender, nationality, or academic age affect review decisions? An analysis of submissions to the journal Biological Conservation", *Biological Conservation*, Vol. 142 No. 11, pp. 2415-2418.

Ranstam, J. (2016), "Multiple p-values and Bonferroni correction", *Osteoarthritis and Cartilage*, Vol. 24 No. 5, pp. 763-764.

Ravikumar, S., Agrahari, A., & Singh, S. N. (2015), "Mapping the intellectual structure of scientometrics: A co-word analysis of the journal Scientometrics (2005–2010)", *Scientometrics*, Vol. 102 No. 1, pp. 929-955.

REF2021 (2020). Guidance. https://www.ref.ac.uk/guidance-and-criteria-on-submissions/guidance/

Rothwell, P. M., & Martyn, C. N. (2000), "Reproducibility of peer review in clinical neuroscience: Is agreement between reviewers any greater than would be expected by chance alone?", *Brain*, Vol. 123 No. 9, pp. 1964-1969.

Sánchez, I. R., Makkonen, T., & Williams, A. M. (2019), "Peer review assessment of originality in tourism journals: critical perspective of key gatekeepers", *Annals of Tourism Research*, Vol. 77 No. 1, pp. 1-11.

Sanchez, T. W. (2020), "The most frequently cited topics in urban planning scholarship", *Urban Science*, Vol. 4 No. 1, pp. 4.

Savin, I., & van den Bergh, J. (2021), "Main topics in EIST during its first decade: A computational-linguistic analysis", *Environmental Innovation and Societal Transitions*, Vol. 41, pp. 10-17.

Serenko, A., & Bontis, N. (2018), "A critical evaluation of expert survey-based journal rankings: The role of personal research interests", *Journal of the Association for Information Science and Technology*, Vol. 69 No. 5, pp. 749-752.

Sivertsen, G. (2018), "The Norwegian model in Norway", *Journal of Data and Information Science*, Vol. 3 No. 4, pp. 2–18.

Smith, B. (2018), "Generalizability in qualitative research: Misunderstandings, opportunities and recommendations for the sport and exercise sciences", *Qualitative Research in Sport, Exercise and Health*, Vol. 10 No. 1, pp. 137-149.

Sohrabi, B., & Iraj, H. (2017), "The effect of keyword repetition in abstract and keyword frequency per journal in predicting citation counts", *Scientometrics*, Vol. 110 No. 1, pp. 243-251.

Thelwall, M., Allen, L., Papas, E. R., Nyakoojo, Z., & Weigert, V. (2021), "Does the use of open, non-anonymous peer review in scholarly publishing introduce bias? Evidence from the F1000Research post-publication open peer review publishing model", *Journal of Information Science*, Vol. 47 No. 6, pp. 809-820.

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*, Vol. 43 No. 2, paper 101094. https://doi.org/10.1016/j.lisr.2021.101094

Thelwall, M. & Sud, P. (2021), "Do new research issues attract more citations? A comparison between 25 Scopus subject categories", *Journal of the Association for Information Science and Technology*, Vol. 72 No. 3, pp. 269-279. https://doi.org/10.1002/asi.24401

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*, Vol. 67 No. 5, pp. 1233–1244. doi:10.1002/asi.23474

Thelwall, M. (2021). *Word association thematic analysis: A social media text exploration strategy*. San Rafael, CA: Morgan & Claypool.

Tomkins, A., Zhang, M., & Heavlin, W. D. (2017), “Reviewer bias in single-versus double-blind peer review”, *Proceedings of the National Academy of Sciences*, Vol. 114 No. 48, pp. 12708-12713.

van Lent, M., Overbeke, J., & Out, H. J. (2014), “Role of editorial and peer review processes in publication bias: analysis of drug trials submitted to eight medical journals”, *PloS One*, Vol. 9 No. 8, pp. e104846.

Vere, J., & Gibson, B. (2021), “Variation amongst hierarchies of evidence”, *Journal of Evaluation in Clinical Practice*, Vol. 27 No. 3, pp. 624-630.

Wang, Q., & Sandström, U. (2015), “Defining the role of cognitive distance in the peer review process with an explorative study of a grant scheme in infection biology”, *Research Evaluation*, Vol. 24 No. 3, pp. 271-281.

Woelert, P., & McKenzie, L. (2018), “Follow the money? How Australian universities replicate national performance-based funding mechanisms”, *Research Evaluation*, Vol. 27 No. 3), 184-195.

Woolston, C. (2021), “'A lot of room for bias': UK funder's data point to uneven playing field”, *Nature*, Vol. 591 No. 7851, pp. 683-685.