Does the perceived quality of interdisciplinary research vary between fields?

Interdisciplinary articles might reasonably be submitted to journals in multiple relevant fields, but would they take a different perspective on their quality? If so, which fields would give an interdisciplinary article the most positive evaluation? Answers to these questions might help authors selecting journals and interdisciplinary research evaluators. In response, we assess the extent to which a published refereed journal article was scored differently by expert reviewers from multiple subject-based Units of Assessment (UoAs). Based on 8,015 UK Scopus-indexed journal articles published 2014-20 and evaluated by multiple UoAs, we found only a 54% agreement rate on a 3-point quality scale for a typical interdisciplinary article. Since we estimated the within-UoA agreement rate to be 86%, this gives strong evidence that quality scores vary more between fields than within fields for interdisciplinary research. We also found some hierarchies between fields, in the sense of UoAs that tended to give higher (or lower) scores for the same article than others. The data is more consistent with a partial hierarchy of quality strictness occurring within UoAs than with UoAs using differing quality criteria, but this cannot be proven. Although the results apply to one country and one type of research evaluation, they give the largest scale evidence yet of systematic field differences in quality evaluations for journal articles. This underlines the importance of choosing relevant and sympathetic journals or evaluation panels to submit to. Interdisciplinary research may also fare worse when evaluated by fields with uniform specific quality standards.
Keywords:

Introduction

Academics often have some control over who evaluates their work when selecting journals to publish in, departments to apply to, or field-based panels for national research evaluation exercises. For interdisciplinary research, these would be critical decisions if the quality of the work was judged very differently between fields. For example, an econophysics paper might be judged to be excellent physics but poor econometrics, so the field of its evaluators would be important. It is therefore useful to understand disciplinary differences in evaluation criteria and outcomes, but most studies of research quality have been theoretical or focused on one discipline.

Research quality has multiple definitions, which vary between contexts and purposes (Langfeldt et al., 2020). Currently the most accepted dimensions of research quality are soundness/rigour, originality, scientific value/impact, and societal value/impact (Aksnes, Langfeldt, & Wouters, 2019; Bonaccorsi, 2018, p. 82; Langfeldt et al., 2020). Communication effectiveness and conforming to field norms (e.g., ethics) are sometimes also specified (Mårtensson et al., 2016). For the UK Research Excellence Framework (REF), research output quality comprises "originality, significance and rigour" (REF, 2019a), with significance encompassing both scientific and societal value, although societal value is also assessed separately in the REF with evidence-based narratives. The meaning of the three terms is operationalised through a five-page explanation, with different criteria for the four broad areas that assess REF submissions (pp. 35-40). For example, "scale, challenge and logistical difficulty" (REF, 2019a, p. 35) is important only for Main Panel A (health and life science-related), whilst "a major expansion of the range and the depth of research and its application"

(REF, 2019a, p. 39) is mentioned only for Main Panel D (mainly arts and humanities). On this basis, a logistically challenging but routine health article, such as an international replication study, might rate a high grade in Main Panel A but a low grade in Main Panel D. The judgements depend on access to "a very broad range of expertise and sufficient time to analyse each output in detail. At best, peer review is not a perfect 'measure', and with the time pressures on some REF panels, maintaining consistency and quality of review is very challenging" (Stern, 2016, p. 14).

Science is sometimes thought to be hierarchical, with some disciplines tending to produce higher quality or more important work than others. Proposals for the best field include theology (DiDonato, 2015), internal medicine (Hamankiewicz, 2016), mathematics (Tomakin, 1989), and metaphysics (Dursun & Taşdemir, 2016), but many might consider climate change research to be essential for survival. More generally, practitioners in some academic research areas, such as physics and life sciences, may believe that their work is "hard science" that is more rigorous than "soft science", such as the social sciences (Simonton, 2018; Nature, 2005). These labels have been argued to be sexist (Light et al., 2022) and misleading because the social sciences are "hard" in the sense of complex and difficult to analyse. Researchers in other fields, such as psychology, sometimes aspire for them to be harder in this sense, however (Uher, 2021; Zagaria et al., 2020). Here, hardness associates with mature fields enjoying a broad consensus on methods and topics. This consensus allows methods to evolve for relatively narrow tasks, increasing perceived rigour. The supposed hardness of a field also varies according to its quantitative characteristics, such as the use of graphs (Smith et al., 2000). This suggests a hierarchy of the rigour aspect of research quality, although it does not relate to its significance and originality dimensions. The hard/soft difference has been argued to disappear for current research topics (Cole, 1983), but this is not true in the sense that more consensus and less diversity is evident for "harder" fields (Fanelli & Glänzel, 2013).

Since there are disciplinary differences in the criteria used to judge the quality of academic research, it seems likely that different outcomes would be common for evaluations of interdisciplinarity research. Assessors may therefore take special precautions against this, especially in large scale formal exercises. In REF evaluations, for example, interdisciplinary articles need to meet the originality and significance quality criteria of at least one of the constituent fields but not all of them (REF, 2019a, p. 35) and there are extra procedures to increase fairness (REF, 2019b). It is not known whether such precautions are effective, however, and whether some fields are generally stricter than others. This study addresses the following questions around these issues with a large collection of peer review scores for articles from multiple disciplines.

- RQ1: Does the perceived quality of an interdisciplinary article depend on the field evaluating it?
- RQ2: Do any fields give higher quality ratings to interdisciplinary research?

Differences in research quality judgements

Scientific quality is routinely judged by peer review, which seems to have become entrenched in academia from the 1970s in the USA in response to the need to be accountable for increased government funding (Baldwin, 2018). Whitley's (2000) theory of different disciplinary organisational structures is useful to highlight general trends in disciplinary differences that can influence peer review. It is a simplification, especially due to changes in external pressures on research and changes in research organisation since it was written

(Borlaug & Langfeldt, 2020; Trowler et al., 2012). Whitley's theory is also designed for academic fields, whereas REF UoAs encompass multiple fields. Its advantage compared to other discipline theories (e.g., hard/soft, pure/applied, rural/urban: Becher & Trowler, 2001) is its focus on reputational work, which is closely related to quality judgements.

Some disciplines have formed a broad consensus on the core dimensions of quality to guard access to key resources, such as centrally controlled expensive equipment. Those like high energy physics (Heidler, 2017) are primarily organised as *conceptually integrated bureaucracies* for this reason (Whitley, 2000) and have a high degree of agreement on the main aspects of research, including reputation and quality judgements. Similarly, hierarchies of evidence in health research, although controversial, are examples of relatively stable and standard sets of methods with broadly agreed differences in evidence strength (Stegenga, 2014). Nevertheless, even in the presence of a broad consensus, expert evaluations have been found to be subjective for tasks such as grant reviewing (Cole, Cole, & Simon, 1981; Erosheva et al., 2021). Human factors such as nepotism and sexism (Wennerås & Wold, 1997), politics (Lillis & Curry, 2015), language (Politzer-Ahles et al., 2020), prestige bias (Tomkins et al., 2017), and attitudes towards the importance of metrics (Langfeldt et al., 2021) are potential causes. Reviewers may be particularly critical of articles that challenge the established paradigm (Siler & Strang, 2017) and judgement mistakes by experts may still be common (e.g., Siler, Lee, & Bero, 2015). Within a review panel, the quality control provided by multiple perspectives that should reduce such biases can be affected by members that defer to the specialism expert or to more senior colleagues (Langfeldt, 2004). Thus, universal agreement should not be expected even in the best case of conceptually integrated bureaucracies.

Disciplines with varied research objects and tasks (technical and strategic task uncertainty) and loose interdependence for research practices and reputation, known as *fragmented adhocracies* (Whitley, 2000) may have little agreement on what constitutes high quality research. Instead, "standards are fairly volatile and can be interpreted differently" (Newig & Rose, 2020), with no competing or accepted paradigms. In such a context, the set of people reasonably able to evaluate an article may be small and it may be evaluated by those without the necessary expertise except in the context of submissions to specialist journals and conferences. Individual researchers may focus on different aspects of a study for its quality (Lamont, 2009). There may also be disagreements on aspects of the field, such as statistical philosophies (Amrhein et al., 2019) and the efficacy of a recommendation (Noon, 2018). Quality judgements in this context can be dependent on the interests of the evaluators (Langfeldt, 2004; Travis & Collins, 1991). This can even be a problem for journals, when editors can be expected to recruit reviewers with relevant expertise (Peters & Ceci, 1982). Thus, widespread disagreement should be expected within fragmented adhocracies.

Research quality can also be perceived fundamentally differently within a single discipline. For example, education researchers have different beliefs about what is most important, such as the production of general theories against the importance of rich contextual interpretations (Moss et al., 2009). Single fields can also harbour different perspectives about what constitutes good practice, even operating within different paradigms until one becomes dominant (Kuhn, 1970). This resolution does not necessarily need to occur, however, especially if the different paradigms can coexist when methodological pluralism is accepted as necessary (Barker & Pistrang, 2005). In such cases, research quality judgment differences between practitioners may be systematic but not large. If paradigms coexist but conflict, as in a *polycentric oligarchy* (Whitley, 2000), then contributions must align with one of the competing schools, accepting that the other schools

would not value them. For example, differences in methods, theories, and traditions can cause "fierce competition" within political science (Bonaccorsi, 2018). In such cases, quality difference judgements between practitioners can be large and even deliberate: "correct" in the understanding of the evaluator but wrong from the perspective of those evaluated (Moran, 1998). The Italian research quality evaluation (VQR) had an arbitration group to identify and resolve cases where reviewers sharply disagreed (Bonaccorsi, 2018).

Interdisciplinary research quality judgements

Inter/multi/transdisciplinary research has been defined in different ways (Arnold et al., 2021; Wagner et al., 2011), such as integrated scholars, methods, theoretical frameworks, and interpretations from two or more fields (Aboelela et al., 2007). Here a looser definition is used: research authored by scholars from multiple disciplines.

At one organisational extreme, an interdisciplinary *field* routinely combines approaches from a range of other fields. For example, library and information science is interdisciplinary in the sense that it incorporates substantial contributions from computer science, business, and education (Chang & Huang, 2012). At the other organisational extreme, a one-off study might draw upon an ad-hoc collection of fields to address a novel problem. Such interdisciplinary research might be submitted for journal/panel evaluation to its parent interdisciplinary field, if there is one, or to any of the component fields. An evaluation may then be overly subjective if the evaluators take the perspective of a single component field. For example, originality is particularly difficult to judge, even within a single field, and this is exacerbated for research outside of the evaluators' expertise (Lamont et al., 2007). In addition, researchers tend to value journals in their own field more (Serenko & Bontis, 2018), so may be suspicious of work published in unrecognised venues. Interdisciplinary research may also target societal benefits that field experts may not recognise (Fontana et al., 2022), undermining value judgements from all fields.

A core problem for evaluating interdisciplinary research is that originality is judged against current paradigms in a field, which a non-expert would not know. Additionally, originality can be sought in different aspects of research. A humanities evaluator might focus on method originality whereas a social scientist might be open to a wider range of types (Guetzkow et al., 2004), including promising novel topics. Something may be judged more original if it is in an area thought to be understudied (Lamont & Guetzkow, 2016). An engineer is likely to classify performance or efficiency improvements to existing systems as originality (Shaheen, 2021). In tourism, incremental additions to the field, including imports from other fields, are considered original but the highest level of originality is a radically new idea (Sánchez et al., 2019). In an interdisciplinary field, subjectivity might also be the result of evaluators prioritising one paradigm (e.g., in polycentric oligarchies) or a lack of knowledge of the topic or methods (e.g., in fragmented adhocracies). Evaluations can also be influenced by the perceived character of the researcher and the evaluator's emotional reaction to the work (Lamont et al., 2007), which would presumably be affected by interpersonal connections within fields. Ideally, evaluators should be drawn from multiple fields with communication between them leading to a shared understanding of the value of the work (Huutoniemi, 2012). Nevertheless, accepting different quality standards inherently increases the uncertainty of interdisciplinary evaluations and so evaluators may need to be encouraged to do this (Langfeldt, 2006).

Are some fields stricter than others?

Comparing the strictness of quality reviewing between fields is complicated by the different criteria used, and the lack of strict guidelines. Moreover, it is impossible to directly compare the quality of two different outputs evaluated with different criteria, even if they are written down. Nevertheless, generating a universal system in which all outputs are scored on the same scale is a fundamental part of national research evaluations. Unless all outputs, people or departments assessed are ranked within their fields, such assessments entail implicit relative quality judgements. In the UK this is achieved with standard generic criteria, such as "world-leading in terms of originality, significance and rigour" (REF, 2021), trusting evaluators, supported by broad guidelines, to operationalise these criteria fairly. A review found no evidence that interdisciplinary research had been disadvantaged in REF2014, but some evidence that institutions had tried to avoid submitting interdisciplinary research in the belief that it might (Arnold et al., 2018). One way of comparing the relative strictness of fields is to analyse ratings given by them to the same interdisciplinary outputs. No previous study seems to have done this, however.

Journals often have detailed criteria for reviewers, which are essentially documents to aid judgements about the quality of the submitted work. They sometimes also publish informal discussions of the type of research that they favour (e.g., Eysenck & Eysenck, 1992). Reviewers presumably use these to supplement their internalized notions of quality from previous reading and reviewing. Although rigour seems to be standard in guidelines, originality and significance may also be required. For example, in the relatively hierarchical field of management, guidelines for more prestigious journals are more likely to stress the importance of a theoretical contribution (Seeber, 2020). Bias and disagreement are common, despite guidelines, because of considerable scope for interpretation (Newton, 2010). To a limited extent, journals may publish detailed specialist evaluation criteria for aspects of studies (particularly methods) that would be comparable between fields, such as suggesting a 60% response rate for surveys (JAMA, 2022). No previous study has compared reviewer instruction strictness between fields.

An indirect and partial way to compare the strictness of fields is to analyse journal rejection percentages on the basis that stricter fields might be expected to have higher rejection rates, other factors being equal. This is unfair because researchers can be expected to broadly understand journal quality criteria in their field and avoid submitting work that is unlikely to be accepted. Moreover, rejection rates seem likely to be higher in fields with greater technical and strategic task uncertainty because the methods and object may not be accepted. Similarly, in fields where there is low technical task uncertainty (Whitley, 2020), such as conceptually integrated bureaucracies, researchers can be expected to reliably judge the quality of their work and target an appropriate journal. Nevertheless, large disciplinary differences in journal acceptance rates (e.g., much higher for health than business: Sugimoto et al., 2013; higher for biomedicine than the social sciences: Björk, 2019) are at least broadly consistent with disciplinary differences in quality standards.

Methods

Data

We used provisional scores in March 2022 for articles submitted to REF2021, excluding articles from the University of Wolverhampton as the raw data. Each journal article record

includes the provisional quality score (0 unclassified, 1* recognised nationally, 2* recognised internationally, 3* internationally excellent, or 4* world-leading) for the article as well as the UoA that evaluated the article. We removed the 318 articles scoring 0 because this score might reflect an author not judged to have contributed sufficiently to the work.

An article submitted to multiple UoAs by different authors would get a separate and apparently independent score in each UoA. Such articles are assumed to be interdisciplinary because they have authors associated with different UoAs. This is an oversimplification because an author might be submitted to the UoA of their department even if it is not a particularly good fit to their research.

Analysis: RQ1

A quarter (25.1%) of the REF journal articles examined had been submitted multiple times by different authors, sometimes in different UoAs. Such articles did not always have the same provisional REF scores. If two scores for an article were selected at random then the chance that they agreed on the four-point scale was 79.8%.

To address the first research question, we first calculated the agreement rate for the scores an article submitted multiple times to the same UoA separately within each UoA. This gave an overall mainly monodisciplinary agreement percentages for each UoA. Calculating separately for each UoA is important because each had d its own procedures for dealing with duplicates. We calculated the within-UoA agreement rate as a baseline to compare the between UoA agreement rate for interdisciplinary research with.

We calculated agreement rates for the same article submitted to different UoAs (i.e., interdisciplinary research) as an estimate of the agreement rates for interdisciplinary research since UoAs did not systematically collaborate to agree scores on multiply submitted articles. Nevertheless, each UoA could ask for input from other UoAs on articles that it deemed interdisciplinary. We reported agreement separately rates by the number of copies of the article submitted (e.g., the percentage agreement for all interdisciplinary articles submitted by four of their authors). We did this because it seemed likely that highly submitted articles would be given a uniformly high score, because many authors had decided that it was one of their best outputs.

Analysis: RQ2

For RQ2, we calculated the average score gain for each pair of UoAs for all articles submitted to both UoAs that they had allocated different scores to. The score gain is the higher UoA score subtract the lower UoA score and it is averaged across all articles for which there was a score disagreement for the two UoAs.

Results

RQ1: The rate of agreement on the quality of interdisciplinary research

Before checking agreement rates for interdisciplinary articles it is useful to investigate agreement rates for monodisciplinary articles as a baseline. An agreement level for interdisciplinary research below that of all research would then give evidence that interdisciplinary affected agreement rates.

If two scores were selected at random for the same article within a single UoA (whether interdisciplinary or not) then there would be a 98.9% chance of the two scores being

the same. For most UoAs, the agreement levels are close to 100% (Table 1), suggesting that the evaluators systematically checked for discrepancies or only assessed duplicates once. The very high agreement rates in other cases (above 95%) suggest that either cross-checking took place, but some differences were allowed to remain or that the cross-checking was imperfect (e.g., due to late scores or late changes from a sub-panel member). Lower agreement rates occurred in UoAs with few duplicates. These panels may not have had a systematic procedure to deal with duplicates because there were too few for this to be important. If this logic is correct, then the largest of these UoAs gives the statistically most reliable estimate of the agreement rate for independently assessed articles in a single field: 86.4% (UoA 15). There is no reason to believe that all UoAs have similar underlying score agreement rate but this nevertheless seems reasonable as a default estimate for the average agreement rate for journal article quality scores, in the absence of special procedure for duplicates.

Table 1. Agreement rates for quality scores given to duplicate copies of the same article within a UoA, irrespective of whether it is interdisciplinary or not.

Unit of Assessment	Non-unique articles in UoA	Average agreement 1v2v3v4
1:Clinical Medicine	1249	98.8%
2:Public Health, Health Services and Primary Care	455	99.2%
3:Allied Health Professions, Dentistry, Nursing and Pharmacy	587	98.6%
4:Psychology, Psychiatry and Neuroscience	845	98.6%
5:Biological Sciences	472	98.7%
6:Agriculture, Food and Veterinary Sciences	119	100.0%
7:Earth Systems and Environmental Sciences	346	100.0%
8:Chemistry	306	99.8%
9:Physics	560	100.0%
10:Mathematical Sciences	384	100.0%
11:Computer Science and Informatics	372	100.0%
12:Engineering	864	97.1%
13:Architecture, Built Environment and Planning	95	100.0%
14:Geography and Environmental Studies	243	95.5%
15:Archaeology	44	86.4%
16:Economics and Econometrics	144	99.3%
17:Business and Management Studies	2112	100.0%
18:Law	116	100.0%
19:Politics and International Studies	152	98.7%
20:Social Work and Social Policy	107	97.2%
21:Sociology	31	100.0%
22:Anthropology and Development Studies	12	100.0%
23:Education	189	99.5%
24:Sport and Exercise Sciences, Leisure and Tourism	284	96.8%
25:Area Studies	1	100.0%
26:Modern Languages and Linguistics	20	100.0%
27:English Language and Literature	12	100.0%
28:History	40	100.0%
30:Philosophy	29	100.0%
32:Art and Design: History, Practice and Theory	24	66.7%
33:Music, Drama, Dance, Performing Arts, Film and Screen Studies	18	94.4%
34:Communication, Cultural and Media Studies, Library & Info Man	22	86.4%

The agreement rates for the same article submitted to *different* UoAs (and therefore interdisciplinary according to the definition here) is 58.9% overall, but higher for articles submitted many times. The linear trend between the agreement rate and number of times submitted can be extrapolated to estimate the theoretical agreement rate for outputs submitted once, if they had accidentally been evaluated by a different but appropriate UoA. This gives a between-UoA projected agreement rate for single submitted outputs of about 53% (Figure 1). Both these figures are substantially lower than the within-UoA agreement rate overall estimated above (86.4%) as well as being substantially lower than the within-UoA agreement rate of each individual UoA (Table 1). Thus, there is strong evidence that there are field differences in the criteria or standards used to evaluate interdisciplinary research, giving a positive answer to RQ1.

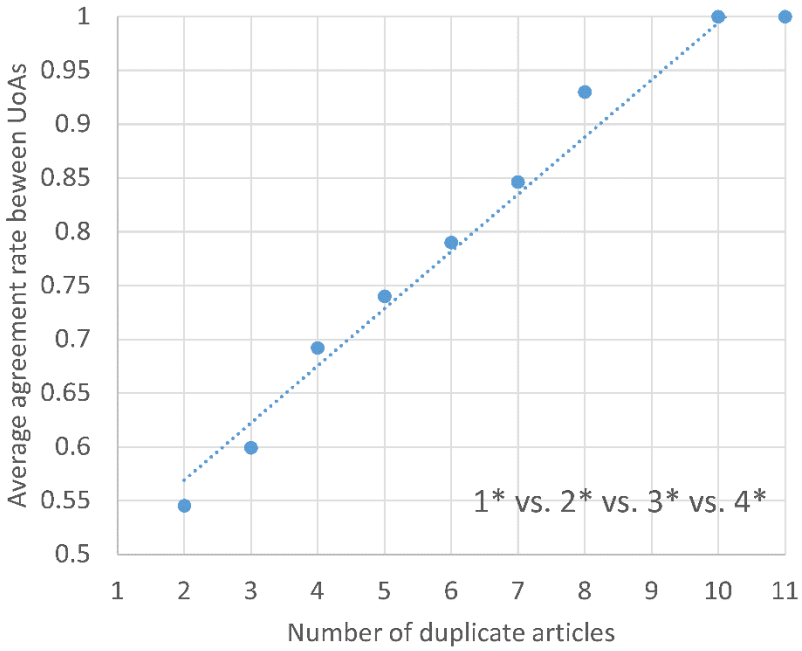

Figure 1. Agreement rate against number of duplicate submissions of the same article scoring at least 1*, only checking scores between different UoAs.

RQ2: Do any fields give higher quality ratings to interdisciplinary research?

Comparing the score given by one UoA to an article with the score given by a different UoA to the same article reveals some trends in terms of UoAs that tend to give higher or lower quality scores, when there is a difference. The results show substantial systematic tendencies in some cases (Table 2). In terms of average scores, the clearest case is UoA 1, which usually gives a lower score to an article than another UoA, when there is a difference: an average 0.56 lower, over 1522 differences. At the other extreme, UoA 24 usually gives a higher score to an article than another UoA, when there is a difference: an average 0.55 higher, over 255 differences. The fact that the average gain for UoA 1 is less than 1 shows that the pattern is not universal and it sometimes gives a lower score than another UoA for the same article. Thus, either UoA 1 has stricter quality standards, or it judges interdisciplinary research by generally stricter standards than the UoAs that shares its interdisciplinary research. In either case, this is clear evidence of different standards between fields, as represented by UoAs.

Table 2. Average score gain by a UoA compared to UoAs for articles that receive different scores between UoAs. The average is across all articles with a score difference.

Unit of Assessment	Articles	Comparisons	Score differences	Average gain
31:Theology and Religious Studies	302	6	2	-1.00
1:Clinical Medicine	11928	4418	1522	-0.56
9:Physics	5473	802	247	-0.42
28:History	1964	40	20	-0.40
19:Politics and International Studies	3064	156	87	-0.33
5:Biological Sciences	7081	2584	958	-0.24
16:Economics and Econometrics	2127	275	137	-0.15
14:Geography and Environmental Studies	4026	969	332	-0.10
15:Archaeology	692	149	57	-0.09
27:English Language and Literature	1474	39	24	-0.08
8:Chemistry	3682	989	291	-0.08
4:Psychology, Psychiatry and Neuroscience	9640	1897	767	-0.08
17:Business and Management Studies	15557	785	417	-0.07
32:Art and Design: History, Practice and Theory	1710	126	25	-0.04
2:Public Health, Health Services and Primary Care	4872	2715	784	-0.01
30:Philosophy	1036	9	2	0.00
21:Sociology	1752	228	139	0.04
18:Law	3384	99	54	0.04
22:Anthropology and Development Studies	1152	111	62	0.06
10:Mathematical Sciences	5819	386	131	0.13
34:Communication, Cultural and Media Studies, Library and Information Management	1382	94	50	0.16
11:Computer Science and Informatics	5547	524	230	0.17
33:Music, Drama, Dance, Performing Arts, Film and Screen Studies	948	39	17	0.18
26:Modern Languages and Linguistics	1563	74	44	0.18
13:Architecture, Built Environment and Planning	2994	293	135	0.20
25:Area Studies	726	63	30	0.20
20:Social Work and Social Policy	3995	520	280	0.22
12:Engineering	17924	1784	741	0.26
7:Earth Systems and Environmental Sciences	4347	1224	425	0.28
23:Education	4062	259	145	0.32
6:Agriculture, Food and Veterinary Sciences	3405	868	333	0.40
3:Allied Health Professions, Dentistry, Nursing and Pharmacy	11376	3364	1334	0.44
24:Sport and Exercise Sciences, Leisure and Tourism	3430	517	255	0.55
29:Classics	225	4	1	1.00

Score differences between UoAs on a pairwise basis reveals which pairs of UoAs are a strict and lenient combination: scores for the same articles consistently lower than the other (Figure 3). The clearest case is again UoA 1: whenever there is a systematic difference for at least 10 (Figure 2) or 30 (Figure 3,4) articles, it tends to give lower quality scores. Conversely, in all cases where there is a systematic difference for at least 10 (Figure 2) or 30 (Figure 4) articles, UoA 24 tends to give higher scores. For other UoAs for which there are many differences, the UoA tends to give higher scores than some UoAs but lower scores than others. For example, UoA 9 tends to give lower scores than UoAs 4, 7, 8, 10, 12 but slightly higher scores than UoA 5.

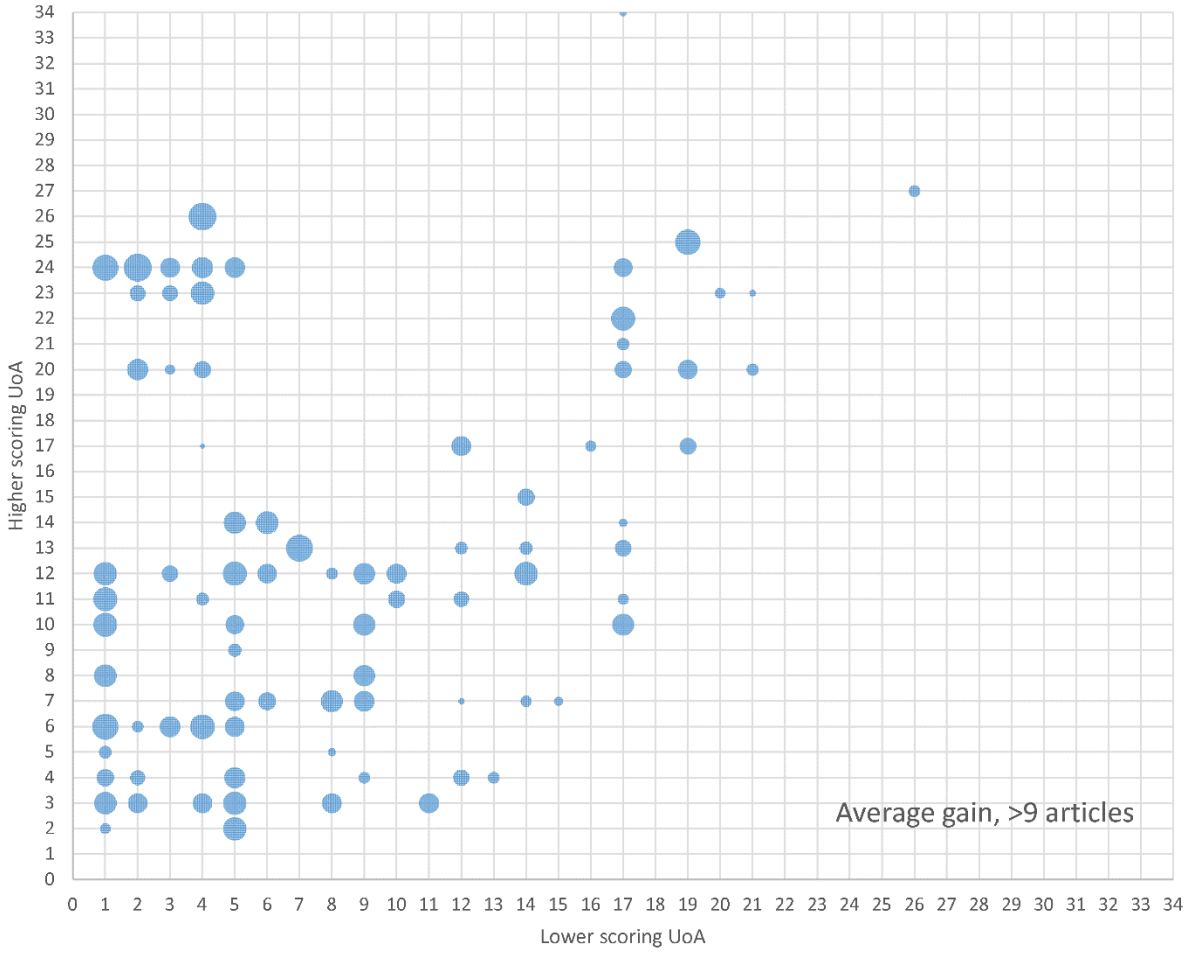

Figure 2. Average score gain by a UoA compared to another UoA for articles that receive different scores in the two UoAs, only plotting pairs of UoAs with score differences for at least 10 articles. The area of each bubble represents the average score gain of the higher scoring HEI relative to the lower scoring HEI, with the highest being 1.1.

Considering only pairs of UoAs for which there are at least 30 shared articles with a score difference (Figure 3, 4), the average gain calculation forms a hierarchical relationship (in mathematical terms, the UoAs form a partially ordered set). In particular, if UoA x tends to score higher than UoA y and UoA y tends to score higher than UoA z then it never happens that UoA x scores lower than UoA z. The same is true for longer chains of UoAs. Thus, the UoAs can be drawn as a partial network, where nodes to the left tend to score articles lower than nodes to the right, when there are at least articles in common (Figure 3). This suggests the existence of a tendency for a hierarchical relationship of quality strictness to prevail in

most or all UoAs. Ignoring UoAs for which there is no comparative information, the strictest UoAs are 1 and 9 and the least strict are 7, 11, 13, 20, 23, 24 in the sense that they never have a lower average score than any other UoA. The same is not true if the number of articles is relaxed to 10, however, perhaps because of the greater role of chance for smaller numbers. For example, in Figure 2: UoA 8 > UoA 5 > UoA 9 > UoA 8, forming a loop.

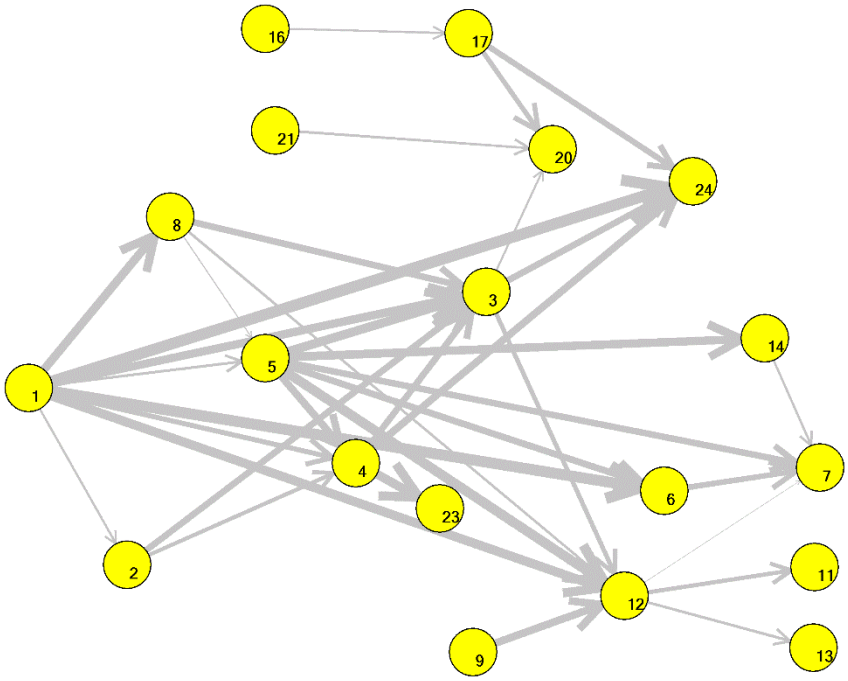

Figure 3. Network illustrating Figure 4. Arrows indicate score gains from the source UoA to the target UoA. UoAs are arranged in a partial left-to-right order for ease of visual interpretation only.

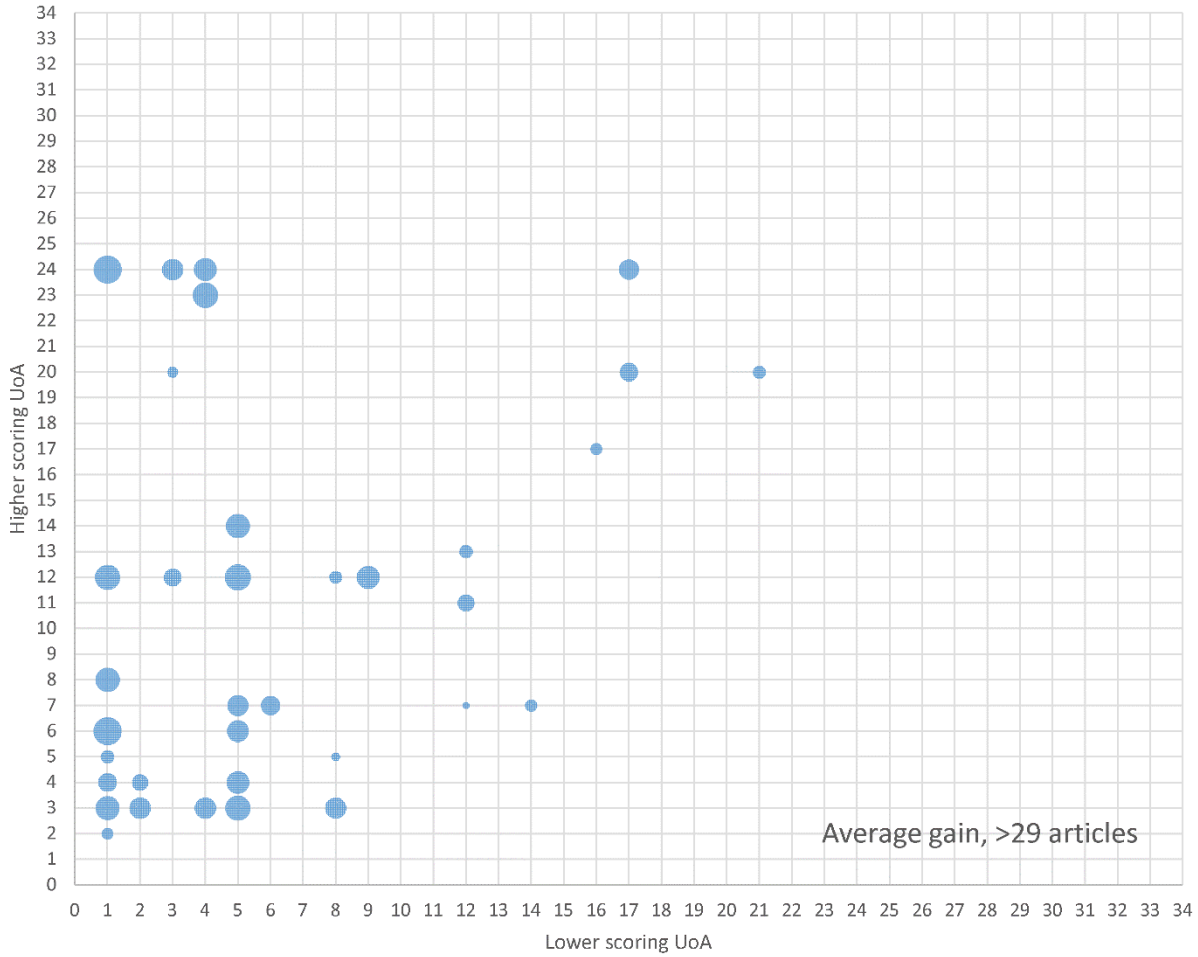

Figure 4. As Figure 2, but only plotting pairs of UoAs with score differences for at least 30 articles. The highest score gain is 0.95.

To check for possible confounding factors at the institutional level, the average score gain was calculated for each UK Higher Education Institution for duplicate articles between UoAs. This was correlated against the average REF score of each HEI's journal articles as an indicator of its prestige. The low result (Pearson's r=-0.007, n=139 HEIs) suggests that institutional prestige could not be a second order explanation for the score differences (e.g., due to more prestigious institutions being given higher scores for the same article and tending to submit to particular UoAs).

Discussion

The results are limited by a focus on one country, one field categorisation scheme, one period and one type of evaluation: a systematic national exercise. They also only apply to collaborative research since an article needs at least two authors to be submitted to two different UoAs. The findings may not apply to journal refereeing, individual researcher evaluation, and evaluations of non-journal outputs. The results also assume that an article submitted to two UoAs is interdisciplinary and relevant to both UoAs but this is not necessarily always true. An article might be monodisciplinary but submitted by at least one researcher employed by an out-of-field department (e.g., an economist in a business school), or the researcher may have changed fields and had to submit one article to an irrelevant UoA.

For RQ1, the results show, apparently for the first time, that agreement rates between fields for interdisciplinary research are lower than agreement rates within fields for the quality of a published journal article. This is based on a weak estimate of within-field agreement rates, however. The result is unsurprising given that refereeing of interdisciplinary outputs is inherently more uncertain than for single field outputs because of the wider range of factors that need to be considered and the judgement needed about whether the quality criteria of one or more fields can safely be ignored (Langfeldt, 2006). This would especially apply if at least one of the fields had agreed quality standards. Even though the REF has substantial procedures for ensuring that interdisciplinary research is evaluated fairly, this does not mean that they can be evaluated with the same degree of certainty as monodisciplinary research. The results can therefore be interpreted either as disciplinary differences in quality evaluations for interdisciplinary research or as greater uncertainty in the evaluations of interdisciplinary research. The former is consistent with a positive answer to the second research question, however. This is concerning given the widespread recognition of the importance of interdisciplinary research for tackling major societal problems (e.g., Alford & Head, 2017; Gibbons, 2000).

For RQ2, the results show, again apparently for the first time, that there is a partial field hierarchy in the strictness of quality evaluations of published journal articles. The evidence of this is that there are systematic trends in which UoA gives a higher quality score to a journal article submitted to two UoAs. This is not an expected finding given that the panel guidelines are not hierarchical but allow fields to specify different quality criteria. The results are broadly consistent with hard sciences being stricter than soft sciences but this pattern is not strong in the data because the hardness of disciplines has not been defined and the pairs of disciplines with enough shared articles to judge differences tend to be similar types (e.g., the same main panel). In addition, Theology and Religious Studies and History, which have never been described as hard sciences, are first and fourth in Table 2.

The partial hierarchies found may well not be due to underlying differences in difficulty or quality judgements. For example, one field might value collaborative types of research or interdisciplinary collaborations more than another, hence scoring such articles higher for their interdisciplinary component rather than for looser quality standards. This could be a second order effect of interdisciplinary research tending to be more applied, thereby scoring higher for significance. Another potential explanation is that one field may have a consensus on core research standards, for example as a conceptually integrated bureaucracy, and be resistant to relaxing any aspect of its core standards for interdisciplinary research. In contrast, the other field might enjoy more flexibility, perhaps as a fragmented adhocracy, and focus instead on how the article might make a valid contribution to knowledge.

Conclusion

The findings confirm that, as believed by many UK academics (Arnold et al., 2018), quality evaluations of interdisciplinary journal articles are more problematic than those of monodisciplinary journal articles. Even if there is no bias against interdisciplinary research, the choice of evaluating discipline can influence the likely score, with some relevant disciplines being apparently stricter than others. Thus, researchers with a degree of choice about where their research can be evaluated, as in the REF, should consider their options carefully in the light of the quality evaluation criteria of the relevant fields. Moreover, the

increased uncertainty of scores for interdisciplinary research may need substantial work to address, given that the REF already had elaborate precautions for this.

The suggestions above also apply to researchers submitting to academic journals on the assumption that similar findings would apply to them. Moreover, authors of interdisciplinary articles should be aware that rejection by journals from one field does not imply that rejection from journals in other fields is also likely.

Finally, the apparent hierarchy of published journal article quality evaluation strictness, if not an accidental by-product of the factors discussed above is an issue of obvious concern to all parties to research evaluation systems that attempt to provide a fair playing field for a variety of disciplines. Whilst research quality is inherently subjective, demonstrable systematic differences suggest a degree of unfairness that needs to be fully identified and rectified.

References

Aboelela, S. W., Larson, E., Bakken, S., Carrasquillo, O., Formicola, A., Glied, S. A., & Gebbie, K. M. (2007) 'Defining interdisciplinary research: Conclusions from a critical review of the literature', *Health Services Research*, 42/1p1, 329-346.

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

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.

Alford, J., & Head, B. W. (2017) 'Wicked and less wicked problems: a typology and a contingency framework', *Policy and Society*, 36/3, pp. 397-413.

Amrhein, V., Greenland, S., & McShane, B. (2019) 'Scientists rise up against statistical significance', *Nature*, 567, pp. 305-307

Arnold, A., Cafer, A., Green, J., Haines, S., Mann, G., & Rosenthal, M. (2021) 'Perspective: Promoting and fostering multidisciplinary research in universities', *Research Policy*, 50/9, 104334.

Arnold, E., Simmonds, P. Farla, K., Kolarz, P., Mahieu, B., & Nielsen, K. (2018) 'Review of the research excellence framework: evidence report'. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/768162/research-excellence-framework-review-evidence-report.pdf

Baldwin, M. (2018) 'Scientific autonomy, public accountability, and the rise of "peer review" in the Cold War United States', *Isis*, 109/3, pp. 538-558.

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

Becher, T., & Trowler, P. (2001). *Academic tribes and territories*. Oxford: McGraw-Hill Education.

Björk, B. C. (2019) 'Acceptance rates of scholarly peer-reviewed journals: a literature survey', *Profesional de la Información*, 28/4, e280407. https://doi.org/10.3145/epi.2019.jul.07

Bonaccorsi, A. (2018) 'Peer review in social sciences and humanities. Addressing the interpretation of quality criteria', In Bonaccorsi, A. (ed.) *The Evaluation of Research in Social Sciences and Humanities,* pp. 71-101. Berlin, Germany: Springer.

Borlaug, S. B., & Langfeldt, L. (2020) 'One model fits all? How centres of excellence affect research organisation and practices in the humanities', *Studies in Higher Education*, 45/8, pp. 1746-1757.

Chang, Y. W., & Huang, M. H. (2012) 'A study of the evolution of interdisciplinarity in library and information science: Using three bibliometric methods', *Journal of the American Society for Information Science and Technology*, 63/1, pp. 22-33.

Cole, S., Cole, J. R., & Simon, G. A. (1981) 'Chance and consensus in peer review', *Science*, 214/4523, pp. 881-886.

Cole, S. (1983) 'The hierarchy of the sciences?', *American Journal of Sociology*, 89/1, pp. 111-139.

DiDonato, N. C. (2015) 'Theology as "queen of science" reconsidered: A basis for scientific realism', *Theology and Science*, 13/4, pp. 409-424.

Dursun, Ş., & Taşdemir, C. (2016) 'Is metaphysics hyper-physics or over-physics? Evaluating it with mathematical paradigms', *Journal of Theoretical Educational Science*, 9/1, pp. 130-145.

Erosheva, E. A., Martinková, P., & Lee, C. J. (2021) 'When zero may not be zero: A cautionary note on the use of inter-rater reliability in evaluating grant peer review', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184/3, pp. 904-919.

Eysenck, H. J., & Eysenck, S. B. G. (1992) 'Peer review: advice to referees and contributors', *Personality and Individual Differences*, 13/4, pp. 393-399.

Fanelli, D., & Glänzel, W. (2013) 'Bibliometric evidence for a hierarchy of the sciences', *PLoS One*, 8/6, e66938.

Fontana, M., Iori, M., Sciabolazza, V. L., & Souza, D. (2022) 'The interdisciplinarity dilemma: public versus private interests', *Research Policy*, 51/7, 104553.

Gibbons, M. (2000) 'Mode 2 society and the emergence of context-sensitive science', *Science and Public Policy*, 27/3, pp. 159-163.

Guetzkow, J., Lamont, M., & Mallard, G. (2004) 'What is Originality in the Humanities and the Social Sciences?', *American Sociological Review*, 69/2, pp. 190-212.

Hamankiewicz, M. (2016) 'Internal medicine: the queen of science', *Polskie Archiwum Medycyny Wewnetrznej*, 126/12, pp. 1050-1053.

Heidler, R. (2017) 'Epistemic cultures in conflict: The case of astronomy and high energy physics', *Minerva*, 55/3, pp. 249-277.

Huutoniemi, K. (2012) 'Communicating and compromising on disciplinary expertise in the peer review of research proposals', *Social Studies of Science*, 42/6, pp. 897-921.

JAMA (2022) 'Instructions for Authors'. https://jamanetwork.com/journals/jama/pages/instructions-for-authors

Kuhn, T. S. (1970). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

Lamont, M., Fournier, M., Guetzkow, J., Mallard, G., & Bernier R. (2007) 'Evaluating creative minds: The assessment of originality in peer review'. In A. Sales, & M. Fournier (Eds.). *Knowledge, Communication and Creativity*, pp. 166–181. London: Sage. https://doi.org/10.4135/9781446215548.n10

Lamont, M., & Guetzkow, J. (2016) 'How quality is recognized by peer review panels: The case of the humanities'. In: Oschner, M. & Hug, S. (eds.) *Research Assessment in the Humanities*, pp. 31-41. Berlin, Germany: Springer.

Lamont, M. (2009). *How professors think*. Cambridge, MA: Harvard University Press.

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, pp. 115-137.

Langfeldt, L., Reymert, I., & Aksnes, D. W. (2021) 'The role of metrics in peer assessments', *Research Evaluation*, 30/1, pp. 112-126.

Langfeldt, L. (2004) 'Expert panels evaluating research: decision-making and sources of bias', *Research Evaluation*, 13/1, pp. 51-62.

Langfeldt, L. (2006) 'The policy challenges of peer review: managing bias, conflict of interests and interdisciplinary assessments', *Research Evaluation*, 15/1, pp. 31-41.

Light, A. E., Benson-Greenwald, T. M., & Diekman, A. B. (2022) 'Gender representation cues labels of hard and soft sciences', *Journal of Experimental Social Psychology*, 98, 104234.

Lillis, T., & Curry, M. J. (2015) 'The politics of English, language and uptake: The case of international academic journal article reviews', *AILA Review*, 28/1, pp. 127-150.

Mårtensson, P., Fors, U., Wallin, S. B., Zander, U., & Nilsson, G. H. (2016) 'Evaluating research: A multidisciplinary approach to assessing research practice and quality', *Research Policy*, 45/3, pp. 593-603.

Moran, G. (1998) *Silencing scientists and scholars in other fields: Power, paradigm controls, peer review, and scholarly communication*. New York, NY: Greenwood Publishing Group.

Moss, P. A., Phillips, D. C., Erickson, F. D., Floden, R. E., Lather, P. A., & Schneider, B. L. (2009) 'Learning from our differences: A dialogue across perspectives on quality in education research', *Educational Researcher*, 38/7, pp. 501-517.

Nature. (2005) 'In praise of soft science', *Nature*, 435/7045, pp. 1003.

Newig, J., & Rose, M. (2020) 'Cumulating evidence in environmental governance, policy and planning research: towards a research reform agenda', *Journal of Environmental Policy & Planning*, 22/5, pp. 667-681.

Newton, D. P. (2010) 'Quality and peer review of research: an adjudicating role for editors', *Accountability in Research*, 17/3, pp. 130-145.

Noon, M. (2018) 'Pointless diversity training: Unconscious bias, new racism and agency', *Work, Employment and Society*, 32/1, pp. 198-209.

Peters, D. P., & Ceci, S. J. (1982) 'Peer-Review Practices of Psychological Journals: The Fate of Published Articles, Submitted Again', *The Behavioral and Brain Sciences*, 5, pp. 187-255.

Politzer-Ahles, S., Girolamo, T., & Ghali, S. (2020) 'Preliminary evidence of linguistic bias in academic reviewing', *Journal of English for Academic Purposes*, 47, 100895.

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

REF (2019b) 'Interdisciplinary Research'. https://www.ref.ac.uk/about-the-ref/interdisciplinary-research/

REF (2021) 'Guide to the REF results'. https://ref.ac.uk/about-the-ref/interdisciplinary-research/

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*, 77, 1-11.

Seeber, M. (2020) 'How do journals of different rank instruct peer reviewers? Reviewer guidelines in the field of management', *Scientometrics*, 122/3, pp. 1387-1405.

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*, 69/5, pp. 749-752.

Shaheen, M. (2021) 'The concept of originality in academic research of engineering', *Education Research International*, article 9462201. https://doi.org/10.1155/2021/9462201

Siler, K., Lee, K., & Bero, L. (2015) 'Measuring the effectiveness of scientific gatekeeping', *Proceedings of the National Academy of Sciences*, 112/2, pp. 360-365.

Siler, K., & Strang, D. (2017) 'Peer review and scholarly originality: Let 1,000 flowers bloom, but don't step on any', *Science, Technology, & Human Values*, 42/1, pp. 29-61.

Simonton, D. K. (2018) 'Hard science, soft science, and pseudoscience: Implications of research on the hierarchy of the sciences'. In: Allison B. Kaufman, James C. Kaufman (eds.) *Pseudoscience: The conspiracy against science*, pp. 77-99, Cambridge, MA: MIT Press.

Smith, L. D., Best, L. A., Stubbs, D. A., Johnston, J., & Archibald, A. B. (2000) 'Scientific graphs and the hierarchy of the sciences: A Latourian survey of inscription practices', *Social Studies of Science*, 30/1, pp. 73-94.

Stegenga, J. (2014) 'Down with the hierarchies', *Topoi*, 33/2, pp. 313-322.

Stern, N. (2016) 'Building on success and learning from experience: An independent review of the Research Excellence Framework'. https://www.gov.uk/government/publications/research-excellence-framework-review

Sugimoto, C. R., Lariviére, V., Ni, C., & Cronin, B. (2013) 'Journal acceptance rates: a cross-disciplinary analysis of variability and relationships with journal measures', *Journal of Informetrics*, 7/4, pp. 897-906.

Tomakin, F. Y. G. (1989) 'Mathematics is the queen of science and the theory of numbers is the queen', *The Philippine Scientist*, 26/1, pp. 60-64.

Tomkins, A., Zhang, M., & Heavlin, W. D. (2017) 'Reviewer bias in single-versus double-blind peer review', *Proceedings of the National Academy of Sciences*, 114/48, pp. 12708-12713.

Travis, G. D. L., & Collins, H. M. (1991) 'New light on old boys: Cognitive and institutional particularism in the peer review system', *Science, Technology, & Human Values*, 16/3, pp. 322-341.

Trowler, P., Saunders, M., & Bamber, V. (Eds.) (2012). *Tribes and Territories in the 21st-Century*. London: Routledge.

Uher, J. (2021) 'Psychology's status as a science: Peculiarities and intrinsic challenges. Moving beyond its current deadlock towards conceptual integration', *Integrative Psychological and Behavioral Science*, 55/1, pp. 212-224.

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., & Börner, K. (2011) 'Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature', *Journal of Informetrics*, 5/1, pp. 14-26.

Wennerås, C., & Wold, A. (1997) 'Nepotism and sexism in peer-review', *Nature*, 387/6631, pp. 341.

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

Whitley, R. (2011) 'Changing governance and authority relations in the public sciences', *Minerva*, 49/4, pp. 359-385.

Zagaria, A., Ando, A., & Zennaro, A. (2020) 'Psychology: A giant with feet of clay', *Integrative Psychological and Behavioral Science*, 54/3, pp. 521-562.