Article:
Thelwall, M., Kousha, K., Stuart, E. et al. (2023) Do bibliometrics introduce gender, institutional or interdisciplinary biases into research evaluations? Research Policy, 52 (8). 104829. ISSN: 0048-7333

https://doi.org/10.1016/j.respol.2023.104829

Research Policy 52 (2023) 104829

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

Do bibliometrics introduce gender, institutional or interdisciplinary biases into research evaluations?

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

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

ARTICLE INFO

Keywords:
Research bias
Gender
Peer review
REF2021
Interdisciplinarity

ABSTRACT

Systematic evaluations of publicly funded research sometimes use bibliometrics alone or bibliometric-informed peer review, but it is not known whether bibliometrics introduce biases when supporting or replacing peer review. This article assesses this by comparing three alternative mechanisms for scoring 73,612 UK Research Excellence Framework (REF) journal articles from all 34 field-based Units of Assessment (UoAs) 2014–17: REF peer review scores, field normalised citations, and journal average field normalised citation impact. The results suggest that in almost all academic fields, bibliometric scoring can disadvantage departments publishing high quality research, as judged by peer review, with the main exception of article citation rates in chemistry. Thus, introducing journal or article level citation information into peer review exercises may have a regression to the mean effect. Bibliometric scoring slightly advantaged women compared to men, but this varied between UoAs and was most evident in the physical sciences, engineering, and social sciences. In contrast, interdisciplinary research gained from bibliometric scoring in about half of the UoAs, but relatively substantially in two. In conclusion, out of the three potential sources of bibliometric bias examined, the most serious seems to be the tendency for bibliometric scores to work against high quality departments, assuming that the peer review scores are correct. This is almost a paradox: although high quality departments tend to get the highest bibliometric scores, bibliometrics conceal the full extent of departmental quality advantages, as judged by peer review. This should be considered when using bibliometrics or bibliometric informed peer review.

1. Introduction

Many countries now employ systematic assessments of publicly funded research institutions to evaluate their performance and/or to allocate performance-based funding (Sivertsen, 2017). These may be carried out primarily by peer review, by peer review informed by bibliometrics, or primarily by bibliometrics, and/or other indicators (Sivertsen, 2017). For example, the UK Research Excellence Framework (REF) is an almost pure peer review exercise with journal impact factors banned but article-level citation rates having a minor role in 11 of its 34 field-based Units of Assessment (UoAs – see the first table for a list and the Data subsection for more details of the REF process). Article citations are typically consulted when the reviewers cannot resolve disagreements about the quality score of a journal article (Wilsdon et al., 2015). In contrast, Sweden allocates funding based on bibliometric and other indicators for transparency, reserving peer review for formative research evaluations (Sivertsen, 2017).

The current trend for national research evaluations is to rely on peer review, with bibliometric data sometimes providing a supporting role (DORA, 2023; EU, 2022; Hicks et al., 2015; Wilsdon et al., 2015), but there are moves towards a greater role for artificial intelligence or other data driven approaches with bibliometrics (e.g., Jisc, 2022; ARC, 2022) and indicator-only exercises are still common (e.g., Belgium, Croatia, Denmark, Estonia, Finland, Norway, Poland, Slovakia, Sweden: Sivertsen, 2023). It is important to assess whether bibliometric indicators would introduce biases in these roles, however. For example, if they have institutional focus, author gender, or study type biases, then they could push an assessment into mistaken and/or unethical outcomes. To illustrate this in the UK REF context, if citation scores favour male authors then their use to inform peer review would tend to nudge reviewers into giving higher scores to male-authored research. Despite this important concern, little is known about the biases introduced by

* Corresponding author.
E-mail addresses: m.thelwall@wlv.ac.uk (M. Thelwall), k.kousha@wlv.ac.uk (K. Kousha), emma.stuart@wlv.ac.uk (E. Stuart), meikomakita@wlv.ac.uk (M. Makita), m.abdoli@wlv.ac.uk (M. Abdoli), pauljwilson@wlv.ac.uk (P. Wilson), j.m.levitt@wlv.ac.uk (J. Levitt).

https://doi.org/10.1016/j.respol.2023.104829
Received 11 December 2022; Received in revised form 23 May 2023; Accepted 5 June 2023
Available online 12 June 2023

bibliometrics into national research assessments, whether as primary data or to support peer review.

This article investigates three important potential biases for bibliometric indicators (replacing or supporting peer review): departmental quality, author gender, and article interdisciplinarity. It investigates each from two perspectives: journal-level and article-level bibliometric data. Since departmental scores are often the most important research assessment outcomes, it is important to know whether there are systematic gains or losses from journal-level or article-level bibliometrics. Gender differences are both an ethical (natural justice) issue and an efficiency concern if half of all researchers are devalued. Finally, interdisciplinary research is both difficult to evaluate and widely encouraged in the belief of its scientific and societal value and so needs special attention (e.g., as given the REF). Although bibliometrics are primarily used to inform peer review, this article does not directly investigate how assessors exploit bibliometric data to aid their judgements but instead identifies the directions of the changes likely if peer review decisions are informed or replaced by bibliometric data. For example, if women score more highly on bibliometrics than on peer review then it would be reasonable to assume that bibliometric-informed peer review would give higher scores to women than would peer review alone. From the institutional perspective, overall score shifts (RQ1) may be more important than biases for individual researchers (RQ2) or outputs (RQ3), which they subsume, but the latter biases are still important because they may systematically disadvantage departments with atypical gender combinations or interdisciplinary research contributions. The following research questions drive this study.

- RQ1: Do grades based on article-level or journal-level citation-based indicators favour high quality departments compared to grades based on peer review in any fields?
- RQ2: Do grades based on article-level or journal-level bibliometrics favour female researchers compared to grades based on peer review in any or all fields?
- RQ3: Do grades based on article-level or journal-level bibliometrics favour interdisciplinary articles compared to grades based on peer review in any or all fields?

2. Background

This review assumes that peer judgements are the best available source of evidence about the quality of academic research, with bibliometrics being imperfect indicators of it. This is problematic because peer review has known biases (Lee et al., 2013) and experts often disagree so their judgements are unreliable (Ancaiani et al., 2015). Nevertheless, peer review is preferred because it can consider more dimensions of research quality (see below) than just citation impact (DORA, 2023; EU, 2022; Hicks et al., 2015; Wilsdon et al., 2015). It also does not have the perverse incentives associated with chasing journal impact factors or high citation counts in evaluation systems that emphasize these (Wilsdon et al., 2015).

2.1. Article-level citation-based indicators and research quality

The core rationale behind using citation counts as an indicator of the value, quality or impact of an academic article is that scientists cite to acknowledge prior influences so that an article's citation count reflects its influence on subsequent research (Merton, 1973). The many flaws of this argument include citations being used for other purposes, including comparisons, and being influenced by social factors (Lyu et al., 2021). In addition, citations may rarely play a core role in less hierarchical subjects (Lin, 2018). More fundamentally, research quality is generally thought to encompass three dimensions: rigour, significance, and originality (Langfeldt et al., 2020). Of these, citations probably reflect significance most and it is not clear that they are good indicators of rigour and originality (Aksnes et al., 2019). Moreover, citations do not reflect societal impacts (van Driel et al., 2007). Thus, even from a theoretical perspective, it seems unlikely that citation counts closely correlate with research quality within any field, unless its three dimensions usually coincide for some reason or if societal impact, rigour, and originality all frequently influence citing behaviours.

From an empirical perspective, several studies have compared scientific impact with peer review quality judgements of academic articles to assess whether citation counts could be a quality indicator. For example, a close to zero correlation was found between citation counts and expert ratings of articles in a medical journal (West and McIlwaine, 2002), between most details of methods reporting (i.e., related to rigour) and citation counts for four psychology journals (Nieminen et al., 2006) and for dementia biomarker studies (Mackinnon et al., 2018). The largest scale article-level comparison used four-level peer review quality rating REF scores in 36 Units of Assessment (UoAs) for 19,580 journal articles from 2008, finding zero or negative Spearman correlations in five UoAs: Theology and Religious Studies (−0.2), Classics; Philosophy (−0.1), Art and Design: History, Practice and Theory; Music, Dance, Drama and Performing Arts (0). The remaining correlations were all positive and at least weak, with the strongest in Biological Sciences, Chemistry and Physics (all 0.6) and Clinical Medicine (0.7) (HEFCE, 2015, Table A3). Similar correlations were found between quality ratings and the field normalised citation indicator FWCI (Field Weighted Citation Impact), although the disciplinary differences were less extreme. Spearman correlation strengths were zero or negative in four UoAs: Classics; Music, Dance, Drama and Performing Arts; English Language and Literature; Theology and Religious Studies (−0.1). The strongest correlations were in Clinical Medicine and Physics (both 0.6) (HEFCE, 2015, Table A8). These results suggest that citation counts, whether field/year normalised or not, are imperfect indicators of journal article quality in most academic fields but their value varies greatly between fields, and they are useless in many arts and humanities fields and some social sciences.

Studies that have correlated average citations with average quality scores aggregated at the departmental level have tended to find positive correlations varying in strength from 0.2 to 0.8 (Abramo et al., 2011; Baccini and De Nicolao, 2016; Franceschet and Costantini, 2011; Pride and Knoth, 2018; Rinia et al., 1998; Van Raan, 2006). Most previous investigations of the relationship between departmental average numbers of citations and RAE/REF scores have also found statistically significant positive correlations, although with disciplinary differences (e.g., Mahdi et al., 2008; Jump, 2015; Traag and Waltman, 2019). These reveal little about article-level correlations, however, since correlation coefficients naturally increase when data is aggregated (van Raan, 2004). No prior study has assessed whether bibliometrics systematically advantage higher quality research units, the focus of RQ1.

2.2. Journal-level citation-based indicators and research quality

The Journal Impact Factor (JIF) and other average citation impact indicators for journals are widely consulted in formal and informal research evaluations. At the informal level, academic appointment committees lacking the time to read the candidates' papers might use JIFs to help make quick decisions about the quality of research described in a CV (McKiernan et al., 2019). Individual researchers may also consult JIFs when deciding where to publish (Beshyah, 2019; Sønderstrup-Andersen and Sønderstrup-Andersen, 2008). More formally, some national evaluation systems reward scholars for publishing in journals meeting a JIF threshold or include JIFs in performance-based funding formulae (Sivertsen, 2017), although many countries construct bespoke stratified lists of journals to assess or reward research (Pölönen et al., 2021).

In research evaluation contexts, JIFs have the advantage of being relatively transparent compared to informal ideas of journal prestige shared within a research community. In fields where citation counts are reasonable indicators of research quality, journals with more citations

per article would tend to publish better articles, so journal citation rate calculations would give (imperfect) indicators of research quality. They may also be better indicators of the quality of an article than the article's citations in some fields (Waltman and Traag, 2020). Moreover, in fields where JIFs are well regarded, competition to publish in higher-JIF journals would form a positive feedback loop (Drivas and Kremmydas, 2020) in which higher JIF journals increasingly monopolise research that the field regards as high quality. Nevertheless, highly original research may tend to be published in journals with lower impact factors (Wang et al., 2017), so the novelty quality dimension may be captured poorly by journal metrics.

The many disadvantages of using JIFs for research evaluation have led to the San Francisco Declaration on Research Assessment (DORA, 2023) campaign against them. JIFs have most of the disadvantages of citation counts, as discussed above. For example, in fields where citations are not indicators of research quality, such as the arts and humanities, they are irrelevant (Fuchs, 2014). They are often inappropriately compared between fields, despite large natural variations in field citation rates. There are many technical problems, such as failure to deal appropriately with the skewed nature of citation counts in most, calculation errors, and discrepancies between the numerator and denominator in calculations that allow journals to game the system by overpublishing citable non-article outputs, such as editorials (Jain et al., 2021; Lei and Sun, 2020; Seglen, 1997; Thelwall and Fairclough, 2015). Thus, despite the simplicity and intuitive appeal of JIF-like calculations, they should be interpreted cautiously.

Empirical research assessing whether academics in a field find JIFs to be credible vary between those that find broad acceptance (implicit in: Currie and Pandher, 2020) or rejection (e.g., Hurtado and Pinzón-Fuchs, 2021; Meese et al., 2017). There are two issues: whether journals in a field can be credibly ranked and whether rankings produced by JIF-like calculations agree with expert rankings. Of course, academics are frequently sceptical about expert-based journal rankings too (Bryce et al., 2020) and different expert rankings may disagree substantially (Meese et al., 2017) so there is no "gold standard" against which bibliometric journal rankings can be compared. The second issue has been repeatedly investigated and the answer varies over time for a field (Walters, 2017). Using the expert-based Australian journal strata, Elsevier's Source Normalised Impact per Paper (SNIP) correlated better than the JIF with human judgement in 27 field-based categories. The SNIP advantage may be its normalisation for field differences that makes it more appropriate in large categories containing multiple fields. In the 26 monodisciplinary broad categories checked, the correlations were close to zero in the arts and humanities (0.2), and weak in the social sciences (0.2–0.4) but stronger elsewhere (0.4–0.8), ignoring the multidisciplinary category (Haddawy et al., 2016). SNIP also correlates better than JIF with expert-based rankings of business and management journals (Mingers and Yang, 2017). Journal h-indexes also correlate moderately with human rankings in some fields (Mingers and Yang, 2017; Serenko and Bontis, 2021), perhaps because they combine quality and quantity components, with larger journals being more recognised. For instance, there are stronger associations between departmental h-indexes and REF scores in Biology (ranging from 0.71 to 0.79), Chemistry (0.71 to 0.83), Physics (0.44 to 0.59) and Sociology (0.53 to 0.62) than with institutional normalised citation impact (ranging from 0.37 to 0.67) (Mryglod et al., 2015).

An analysis of the correlation between peer review quality ratings and field/year normalised journal citation rates (SNIP) for 19,130 articles from REF2014 in 36 UoAs published in 2008 found Spearman correlation strengths being zero or negative in four UoAs: Classics (−0.8); Art and Design: History, Practice and Theory; Theology and Religious Studies (−0.1), Arts Area Studies (0). The strongest remaining correlations occurred for Clinical Medicine, Chemistry (all 0.5) and Biological Sciences (0.6) and Economics and Econometrics (0.7) (HEFCE, 2015, Table A18). Thus, as for citation counts, field/year normalised journal impact is an imperfect indicator of journal article quality in most academic fields, its value varies greatly between fields, and it is useless in some arts and humanities fields.

An investigation into the Italian VQR (Valutazione della Qualita della Ricerca) research evaluation 2004–2010 has combined journal impact and citation count data, comparing it with peer review scores from two or three experts, using a four point scale. It analysed 590 economics, management and statistics (Area 13) journal articles for which the VQR process produced both bibliometric and peer review scores. The bibliometric method used a combination of article and journal citation rates (see below) and the peer review method used two independent reviewers, who may have been influenced by bibliometrics (especially since they were known to be important for the VQR). The peer review and bibliometric approaches agreed only moderately (weighted Cohen's kappa of 0.54), but at a higher rate than for the agreement between two independent reviewers (0.40). There was a suggestion of disciplinary differences in the results (Bertocchi et al., 2015).

2.3. Gender bias in academia, peer review and bibliometrics

There is wide suspicion that sexism affects evaluations of the work of female academics because sexism is not yet eradicated from society and because women are underrepresented globally in senior roles (UNESCO, 2022) and for academic prizes (Meho, 2021). Many lists of highly cited scholars are also male dominated. For the Italian VQR research evaluation 2004–2010, outputs (of all types) submitted by women were less likely to receive the top score from post-publication peer review (53 % of the sample) or bibliometrics (47 % of the sample, see below for methods) than research submitted by men, even after accounting for age, seniority, and compulsory maternity leave. This result was not affected by reviewer genders (Jappelli et al., 2017). Nevertheless, the extent of the impact of sexism on peer review scores and citation counts in academia is contested. There are many studies showing that female candidates are or are not discriminated against in evaluations of their research, with no clear outcome (Begeny et al., 2020; Ceci and Williams, 2011). Moreover, overall career statistics and perhaps also prizes favour men because of shorter female career lengths (Huang et al., 2020). It is therefore possible that female-authored research is generally fairly judged in some fields but not others, such as those generating "chilly climates" for women (Biggs et al., 2018; Else, 2018; Overholtzer and Jalbert, 2021). Intersectional factors may well also be relevant, with women that are also from other disadvantaged groups being particularly affected in some or all fields (Banda, 2020; Wilkins-Yel et al., 2019).

Many studies have investigated whether female-authored papers tend to be less cited than male-authored papers, with the suspicion of direct citation sexism through men preferring to cite male authors in some or all fields (e.g., Wang et al., 2021). Sexist citation practices may also be indirect, if the achievements of male authors are more celebrated, making their work more likely to be noticed and cited (Merton, 1968). Similarly, if men tend to cite their friends and these are men then this would generate a second order sexist citation bias against women. The empirical evidence for sexist citation is mixed, however, with the largest-scale evaluation with the most robust citation indicator suggesting a small female citation advantage in six out of seven large predominantly English-speaking countries (Thelwall, 2020). These national averages may hide individual fields where females are slightly less cited, however (e.g., Andersen et al., 2019; Maliniak et al., 2013). Moreover, since female first authored research attracts disproportionately many readers than citers, citations might still systematically underestimate its value (Thelwall, 2018).

The strongest easily evidenced gender difference in academia is between fields rather than in citations. In many countries women numerically dominate some fields (e.g., nursing, allied health professions, veterinary science) and men numerically dominate others (e.g., mathematics, philosophy, physics, engineering) in terms of personnel (UNESCO, 2022) and publications (Thelwall et al., 2019; Thelwall et al.,

2020). Africa may have the least gender variation between fields (at least in terms of students: UNESCO, 2022) and there is paradoxically greater gender inequality between fields in countries where there is less gender inequality overall (Stoet and Geary, 2018; Thelwall and Mas-Bleda, 2020). There is also a male/female gender differentiation within fields, with women more likely to engage in people-related topics, to use qualitative methods (Thelwall et al., 2019; Thelwall et al., 2020) and to have societal progress goals (Zhang et al., 2021). These factors all cause second order effects in bibliometric studies and perhaps also for peer review. Second order gender effects in bibliometrics are likely to occur because topics and fields have different citation rates. Thus, even for a set of researchers within a field, if one gender is more cited than the other then this could be because of differing research methods or specialties rather than sexism affecting choices of citations (e.g., Downes and Lancaster, 2019). It is impossible to fully differentiate between the two because all research is different and narrowing down to a specific enough topic to avoid the likelihood of topic or methods differences is likely to generate too few articles to statistically identify any gender difference, given that it is likely to be small (i.e., large samples are needed to detect small effect sizes).

One study has compared bibliometric scores with post-publication peer review scores for 7500 outputs assessed in the 2010–2014 Italian VQR (Jappelli et al., 2017). For the peer review component, 14 field-based panels of about 30 experts sent each output to two external reviewers (three, when there was a discrepancy) and adjudicated on the results to give a four-point score. Reviewers were asked to check three quality dimensions: originality/innovation, relevance, and internationalisation. Bibliometrics were used for most journal articles (not other output types) in areas primarily producing English-language journal articles: natural and life sciences, engineering, mathematics and statistics, computer science, and economics. The bibliometric scoring system used a combination of citations and journal impact factors. Essentially, each output was given one of four scores by comparing its citation counts to three citation-based thresholds for its field and the same for its journal impact factor. A 4×4 matrix of outcomes was then used to judge what score to assign or whether to apply peer review instead. For example, if both methods got the same score, then it was used but if there was a large disagreement then peer review was called on instead (Ancaiani et al., 2015). The double-assessed sample of 7500 journal articles by design excluded articles where the two bibliometric indicators disagreed (e.g., low cited articles in high impact journals). Separate regressions for the peer review and bibliometric scores found that articles submitted by women were more disadvantaged by peer review than by bibliometrics, even after accounting for age, academic rank and co-authorship (Jappelli et al., 2017). This suggests that it was easier for an Italian woman, compared to an Italian man, to get a highly cited article in a high impact factor journal than for her article to be judged to be excellent by two reviewers (of any gender). This study did not reveal field differences, or the effect of journal impact or citation counts independently, however. A prior UK white paper also suggested that bibliometrics advantaged women compared to peer review in some fields but did not give details (HEFCE, 2015).

2.4. Difficulties evaluating interdisciplinary research

The term "discipline" is sometimes used to denote a research field (e.g., common topics and/or methods) and sometimes to denote a mature field backed by journals, conferences, and departments (Sugimoto and Weingart, 2015). The latter sense is used here. Depending on how it is defined, interdisciplinary research combines theories, methods and/or personnel from multiple disciplines to address a common goal (Aboelela et al., 2007; Arnold et al., 2021; Wagner et al., 2011). For the REF, interdisciplinary research is effectively defined as research that needs the expertise of multiple UoA panels to evaluate (REF, 2019), and this is now it is operationalised in the current article. This does not directly match existing definitions of interdisciplinarity because it is evaluation-focused rather than input- or goal-focused, but it seems likely to have a large overlap in practice because both definition types involve multiple disciplines.

Interdisciplinary research is useful for applied research to address societal issues and for basic science that targets such issues as a longer-term goal (Gibbons et al., 1994; Stokes, 1997). Citation counts are likely to be less useful for evaluating interdisciplinary research than for single discipline research because its significance is more likely to be at least partly determined by non-academics judging its societal value (Gibbons et al., 1994; Whitley, 2000). Thus, factors unrelated to citations seem likely to be more important for interdisciplinary research quality judgements, and it is intrinsically complex to evaluate (Huutoniemi, 2010). There is no large-scale citation-based empirical evidence to support this claim, however.

Citation analyses of interdisciplinary research have tended to evaluate the extent to which average citation counts for interdisciplinary research relate to the average citation counts of the constituent fields. It has been shown, for example, that interdisciplinary research citation counts can tend to be greater or less than the average of the constituent fields, depending on the fields in question (Levitt and Thelwall, 2008). Using three dimensions of diversity (Stirling, 2007), combining a greater number of fields associates with more citations but combining dissimilar fields associates with fewer citations (Yegros-Yegros et al., 2015). Thus, from a citation analysis perspective the effect of interdisciplinarity is unclear and no hypotheses are implied for the overall relationship between interdisciplinarity and research quality.

3. Methods

The research design was to apply bibliometrics to a set of articles with REF2021 peer review scores and assess whether replacing peer review scores with bibliometric equivalents would introduce systematic score shifts suggestive of bias relative to peer review.

3.1. Data from REF2021

The data used in this analysis started from 148,977 confidential provisional REF2021 scores from March 2021 for journal articles submitted by UK academics for assessment. These had to be first published between 2014 and 2020 to be in scope (although articles 2018–2020 and were discarded for the analyses, as described below). Each researcher could submit their best 1–5 outputs, with an average of 2.5 per full time equivalent member of staff, but only the journal articles are considered here. Articles from the University of Wolverhampton were redacted for confidentiality reasons. The 148,977 scores are considered sensitive and had to be deleted by 9 May 2021, with only aggregates being subsequently published. This is a significant dataset because it is the largest ever systematic science-wide post-publication peer review research quality scoring exercise for journal articles. It is over twice the size of a comparable dataset from Italy 2004–10 that included 135,907 journal articles since a minority were evaluated with peer review (Ancaiani et al., 2015). It also has a finer-grained field classification system (34 rather than 14). The scoring is taken seriously because the results are expected to direct about £14 billion of research funding over seven years, and 60 % of this is directly tied to output scores. The size of the dataset is important because it allows fine-grained analyses of fields and differences between them.

REF2021 is split into 34 Units of Assessment (UoAs), each of which has a "subpanel" of expert reviewers, predominantly senior UK academic researchers (over 1000 altogether). Academics must submit all their outputs to a single UoA, normally as part of a departmental submission. As a convenient terminological simplification in the current article, an institution's submissions to a single UoA will be assumed to be a department, although they are likely to often combine departments, be sub-units within larger structures or include a few out-of-department scholars. Each article is allocated at least two primary reviewers from

the UoA subpanel (not external experts in contrast to the VQR) who undertake to read it and agree a single quality score to encompass rigour, originality, and significance, following written field-sensitive guidelines (REF, 2020). The scores are 0 (unrated), 1* (recognised nationally), 2* (recognised internationally), 3* (internationally excellent), or 4* (world leading). As mentioned above, to resolve disagreements for individual articles, assessors in 11 UoAs occasionally consulted article citation counts (comparing them to international field benchmarks for different percentiles from the Web of Science) but bibliometric data was otherwise banned.

The 318 unrated articles were removed because these sometimes indicated that an authorship claim had not been accepted rather than that the article was not of national quality. Many articles were submitted by multiple authors. Such duplicates were removed within UoAs or Main Panels (groups of UoAs), as appropriate for the aggregation level reported. When duplicates had different quality scores, the median was used, or a randomly selected median when there were two. Submitting institutions are only told the percentage of outputs that achieved each score within each UoA and not the scores of individual outputs: these individual scores (as analysed in the current article) were destroyed.

Since the bibliometrics (article level citation counts) were occasionally used in 11 UoAs to arbitrate when two reviewers disagreed on an article, ideally the articles where this applied would be removed. They were not recorded, however, so it is not possible to completely remove the influence of bibliometrics on the peer review scores. Keeping these should only have a minor influence on the results: primarily reducing the apparent biases in a switch from peer review to bibliometrics.

Articles were grouped into Higher Education Institutions (HEIs: universities or research institutions like the Institute of Cancer Research) according to the HEI that had submitted them. The gender analysis uses the first author gender of each article, as recorded in Scopus. Although some fields, such as economics, partially use alphabetical authoring, and corresponding authors are often important, the first author is most likely to be the main contributor in all broad fields (Larivière et al., 2016). When this assumption is wrong and the main contributor has a different gender to the first author then this adds noise to the data, reducing the magnitude of any overall differences found. First authors were assumed to be male or female if their first name, if recorded in Scopus, matched a first name that is used at least 90 % for one gender in the UK, according to GenderAPI.com social media profiles or 1990 US census data (Larivière et al., 2013). Articles with authors having relatively gender-neutral names, such as Sam, were ignored for the gender analysis. Nonbinary genders were not detected because there was no practical non-intrusive way to identify them.

Articles were regarded as interdisciplinary if they were flagged as such in the REF database, either by the submitting institution or the UoA panel members. In the REF context, “interdisciplinary” means that the article may need the expertise from a different UoA to assess, for example, because it is an article about classical physics submitted to the Classics UoA but needing some physics knowledge to understand. Since UoAs seem to encompass one or more disciplines, “interdisciplinary” articles are likely to be interdisciplinary, but the other articles can still be interdisciplinary because multiple disciplines exist within a single UoA. The interdisciplinary flags were not assigned systematically and there were differences between HEIs in the extent to which these labels were used, so the quality of this data is weak. This will reduce the size of any differences between interdisciplinary and other research found in the analysis.

3.2. Citation data from Scopus

For the bibliometric data analysed here (but not available to REF assessors), the articles were matched against a copy of Scopus downloaded in January 2020 (when the bibliometric data for REF2021 would have been available). Articles were primarily matched with Scopus records by Digital Object Identifier (DOI, $n = 133{,}218$), with a few extra ($n = 997$) found by automatically comparing titles and manually checking the results. Only articles from 2014 to 17 were used ($n = 73{,}612$, see also the first table below for exact numbers for the different experiments) to allow at least a three-year citation window (i.e., each article had at least three full years to attract citations), which should give a moderate correlation with long term citations in most fields (Wang, 2013).

Raw citation counts are not useful for the dataset because each UoA combines multiple years and fields (especially for interdisciplinary research). A field and year normalised log-transformed citation score (NLCS) was therefore calculated for each article as follows (Thelwall, 2017). First, all citations were replaced with the log transformation ln(1 + x) to reduce skewing, diminishing the influence of individual highly-cited articles (otherwise all articles in a field could be excessively penalised in the normalisation stage by the presence of one or a few highly cited articles). Next, each log-transformed citation count was divided by the average for the narrow Scopus field and year it was in, giving the NLCS. Articles in multiple fields were instead divided by the average of the relevant field averages (averaging across all articles, not just UK articles). By design, NLCS are unbiased in the sense that they can fairly be compared between articles from different fields and years. An NLCS of 1 is always a world average score and higher values indicate more (log-transformed) citations than average for the field(s) and year of the article. There are up to 330 Scopus narrow fields in each year, so the field normalisation is relatively fine grained.

A limitation of the above approach is that the field classification of Scopus is journal-based and relatively crude (Boyack and Klavans, 2010), undermining the accuracy of the field normalisation. This will tend to add noise to the data and reduce the strength of any differences found.

A journal impact indicator was also calculated for each journal as the mean of the NLCS of all articles in the journal for the given year. This is called here the journal mean NLCS, or JMNLCS. This is an average citation impact indicator for a journal. It is preferable to the well-known Journal Impact Factor because it is field normalised, adjusts for the skewed nature of citation counts (de Solla Price, 1976), and has a longer citation window.

3.3. Analysis

A bibliometric calculation was carried out to mimic the REF procedure to assess how a greater role for bibliometrics (replacing or more systematically informing peer review) might impact the results for departments, women, and interdisciplinary research. The analysis is based on completely replacing all peer review scores with bibliometrics (either article citation rates or journal citation rates) but the direction of any change also points to the influence of bibliometrics if they are used to inform peer review without completely replacing it.

3.3.1. Simulated REF2021 scores from bibliometrics

For each HEI and UoA, the UK REF peer review results are published as the number (in fact the percentage) of outputs rated 1*, 2*, 3*, or 4*. To closely mimic this process with bibliometric data, the articles from each UoA were ranked in order from the lowest to the highest scoring on the NLCS field normalised article-level citation rate bibliometric, and then thresholds set so that the correct number of articles were in each category (HEFCE, 2015; Traag and Waltman, 2019). For example, if a UoA had 5 % 1*, 10 % 2*, 45 % 3* and 40 % 4* in REF2021 then three NLCS thresholds were set (c_1, c_2, c_3) so that 5 % of the articles from that UoA had $NLCS < c_1$, 10 % had $c_1 \leq NLCS < c_2$, 45 % had $c_2 \leq NLCS < c_3$ and 40 % had $c_3 \leq NLCS$. In the case of ties, articles were arranged randomly so that the number of articles in each category was exact.

The above process gives a predicted set of four level quality scores for each UoA based on article citations with a distribution identical to the

REF2021 distribution based on peer review. These will be called the *article citation rate quality scores* hereafter. The same procedure was repeated independently for the JMNLCS field normalised journal-level citation rate indicator to give a simulated set of *journal citation rate quality scores* for each UoA based on journal citation rates.

3.3.2. Departmental score gains from bibliometrics

Within each UoA, the average score of each HEI (i.e., a "department", as introduced above) was calculated by averaging (a) the peer review scores (as in REF2021, but for journal articles only), (b) the *article citation rate quality scores* derived as above, and (c) the *journal citation rate quality scores* derived as above. For example, if a department had 200×1* scores and 100×4* scores (under any of the three methods) in a UoA then its average would be $(200 \times 1 + 100 \times 4)/(200 + 100) = 2$. This calculation is sometimes informally called a department's Grade Point Average (GPA). For each department, the peer review GPA was subtracted from the NLCS GPA to give the GPA increase (possibly negative) expected from a complete switch to scores based on NLCS article level citations. By design, some GPA increases would be positive and others negative, with the departmental size weighted average being zero. This was repeated independently for the JMNLCS journal citation rate scores.

Within each UoA, departmental (i.e., institutional) peer review GPAs were then correlated with GPA NLCS increases to assess for departmental quality bias as measured by article level citation GPA deviation from peer review GPA. A positive correlation would indicate that high peer review GPA departments gained from article level citation scoring, and a negative correlation would indicate that they lost from it, whereas a zero correlation would indicate the lack of a (linear) bias. This correlation is not assessing the accuracy of the predictions but systematic factors behind a switch from peer review to article level citation scoring, or increasing influence for article level citation scoring to support peer review. This was repeated independently for the JMNLCS journal citation rate scores.

3.3.3. Gender and interdisciplinary research score gains from bibliometrics

For the gender analysis, departments were ignored and instead the REF2021 peer review score was subtracted from the *article citation rate quality score* for each article, giving a set of score shifts for males and females. The average female subtract male quality difference was calculated and a 95 % confidence interval for the difference between two means using the t-distribution formula. It would have been simpler to calculate the female citation advantage alone, assuming that the male citation advantage was the opposite, but this assumption would be unsafe because the unknown gender articles might be disproportionately non-British researchers that may have a different relationship between citations and research quality (e.g., if publishing in a language underrepresented in Scopus). This process was repeated for the journal citation rate quality scores.

The above process was repeated for research flagged as interdisciplinary in the REF database, compared to research without the interdisciplinary flag.

4. Results

The number of articles 2014–17 and HEIs varies considerably

Table 1
The number of articles, HEIs, first author male/female genders, and interdisciplinary articles analysed. All were submitted to REF2021 and matching a Scopus journal article 2014–17.

#	UoA or main panel	HEIs	Female	Male	Interdisc	Monodisc	Articles
1	Clinical Medicine	31	1948	2695	682	5289	5971
2	Public Health, Health Services & Primary Care	33	936	984	336	2043	2379
3	Allied Health Professions, Dentistry, Nursing & Pharmacy	89	2296	2124	850	5031	5881
4	Psychology, Psychiatry & Neuroscience	92	1997	2134	499	4496	4995
5	Biological Sciences	44	1252	1841	305	3535	3840
6	Agriculture, Food & Veterinary Sciences	25	610	686	203	1621	1824
7	Earth Systems & Environmental Sciences	40	519	1091	342	1936	2278
8	Chemistry	41	458	1088	294	1817	2111
9	Physics	44	282	1218	229	2913	3142
10	Mathematical Sciences	54	354	1909	314	2819	3133
11	Computer Science & Informatics	89	335	1615	423	2406	2829
12	Engineering	88	1128	4670	1330	9408	10,738
13	Architecture, Built Environment & Planning	37	321	745	99	1345	1444
14	Geography & Environmental Studies	56	561	1005	105	1844	1949
15	Archaeology	24	105	139	22	276	298
16	Economics & Econometrics	25	140	718	25	997	1022
17	Business & Management Studies	107	1753	3771	361	6487	6848
18	Law	67	391	501	59	933	992
19	Politics & International Studies	56	422	856	84	1354	1438
20	Social Work & Social Policy	75	892	722	168	1674	1842
21	Sociology	37	376	380	142	708	850
22	Anthropology & Development Studies	22	193	231	29	489	518
23	Education	82	881	730	202	1670	1872
24	Sport & Exercise Sciences, Leisure & Tourism	60	404	862	154	1403	1557
25	Area Studies	20	107	120	49	214	263
26	Modern Languages & Linguistics	41	263	196	51	487	538
27	English Language & Literature	79	206	162	61	352	413
28	History	76	212	330	44	549	593
29	Classics	17	19	34	9	48	57
30	Philosophy	35	92	333	18	469	487
31	Theology & Religious Studies	22	23	57	16	74	90
32	Art & Design: History, Practice & Theory	69	225	230	58	530	588
33	Music, Drama, Dance, Performing Arts, Film & Screen Studies	68	120	148	46	258	304
34	Communication, Cultural & Media Studies, Library & Information Man.	54	220	244	57	471	528
A	Main Panel A (UoAs 1–6)	128	8168	9350	2519	19,951	22,470
B	Main Panel B (UoAs 7–12)	105	2973	11,260	2778	20,748	23,526
C	Main Panel C (UoAs 13–24)	126	6294	10,379	1395	18,755	20,150
D	Main Panel D (UoAs 25–34)	129	1477	1840	402	3430	3832

between UoAs (Table 1). Overall, the number of articles and HEIs per UoA varies greatly, interdisciplinary research is rare, and men dominate Main Panel B UoAs 7–12.

4.1. RQ1: departmental quality

The first research question asks whether grades from article-level or journal-level citation-based indicators favour high quality departments compared to grades from peer review in any fields. With one exception, HEIs with lower GPAs tend to gain from the bibliometric predictions (Fig. 1). They always gain from JMNLCS predictions and usually gain from NLCS predictions, except in UoA 9. This tendency is moderate or strong in all UoAs except two (1 and 9). The correlations are statistically significantly different from 0 in all cases except UoA 1 Clinical Medicine (NLCS and JMNLCS), UoA 9 Physics (NLCS), and UoA 25 Area Studies (NLCS). Thus, journal-level and article-level bibliometrics disadvantage higher scoring departments in almost all fields.

For the article-level NLCS, the most likely factor behind the result for lower numbered UoAs is that citations imperfectly reflect research quality so that weaker articles occasionally become highly cited, whereas stronger articles sometimes attract few citations. A department consistently producing high quality research could therefore expect to have some rarely cited articles and a department constantly producing lower quality research could expect to have some highly cited articles. Thus, whilst generally higher scoring departments tend to produce more cited research, they tend to be more consistent at producing high quality research than highly cited research. Chemistry is an exception. In this case there is a very weak tendency for the highest scoring departments to be more consistent at producing highly cited work than high quality work. It is possible that some departments selected articles partly on bibliometrics, in the knowledge that they may be consulted in REF evaluations. For higher numbered UoAs, where citations and impact factors have little correlation with research quality, the negative correlations are a statistical effect of replacing genuine scores with almost random noise and then averaging both. Thus, although the magnitudes of the correlations are similar across all UoAs and the practical implications are the same (bias against higher scoring departments), there are at least two distinct causes.

For the journal-level JMNLCS, the above argument largely applies, except for the chemistry exception. Again, for lower numbered UoAs, departments tending to produce high quality research tend to be more consistent in producing high quality articles than in getting them published in high impact journals. Clinical Medicine is a partial exception, in that good departments are almost equally able to produce consistently high-quality research and publish in consistently high impact journals.

4.2. RQ2: gender differences

The second research question asks whether grades based on (article-level or journal-level) bibliometrics favoured female researchers compared to grades based on peer review in any or all fields. There is some evidence of a weak tendency for female first-authored research to gain from bibliometric score allocation in some fields (Fig. 2). The error bars include zero in almost all UoAs and the difference is marginal for the exceptions. It is not reasonable to draw strong conclusions for individual UoAs in this case because the marginal results are to be expected whenever many confidence intervals are drawn, even if there are no underlying differences (Rubin, 2017). Nevertheless, the female advantage is positive in 26 out of 34 UoAs for NLCS ($p = 0.001$ for a post-hoc binomial test for gender difference $\alpha = 0.5$) and in 25 out of 34 UoAs for JMNLCS ($p = 0.001$ for a post-hoc binomial test for gender difference $\alpha = 0.5$), giving statistical evidence of an overall female gain from bibliometrics. Moreover, the difference is statistically significant and positive for journal impact (JMLNLCS) in Main Panel B (mainly physical sciences and engineering). It is also statistically significant and positive for article citations (NLCS) in Main Panel C (mainly social sciences).

4.3. RQ3: interdisciplinary research

The third research question asks whether grades based on (article-level or journal-level) bibliometrics favour interdisciplinary research compared to grades based on peer review in any or all fields. There is some evidence of a moderate tendency for interdisciplinary research to gain from bibliometric score allocation in some fields, but not overall

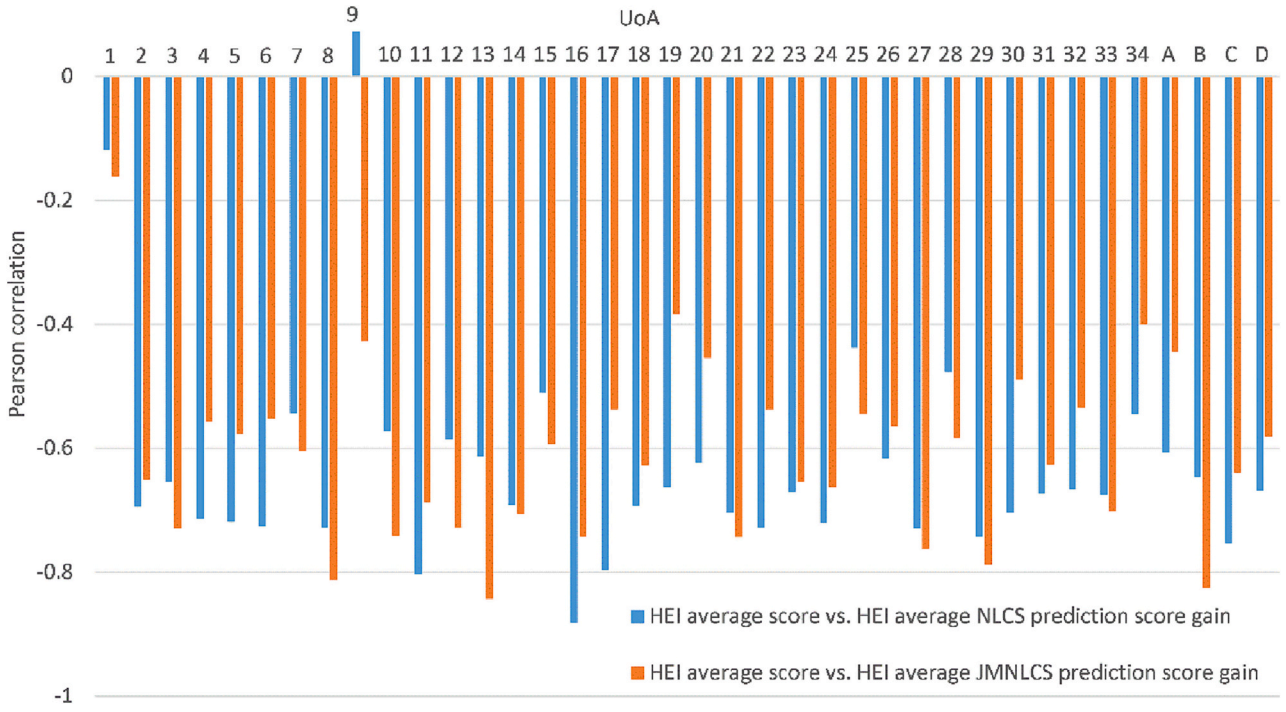

Fig. 1. Pearson correlations between HEI average scores and HEI average prediction gains from allocating scores only with NLCS or JMNLCS.

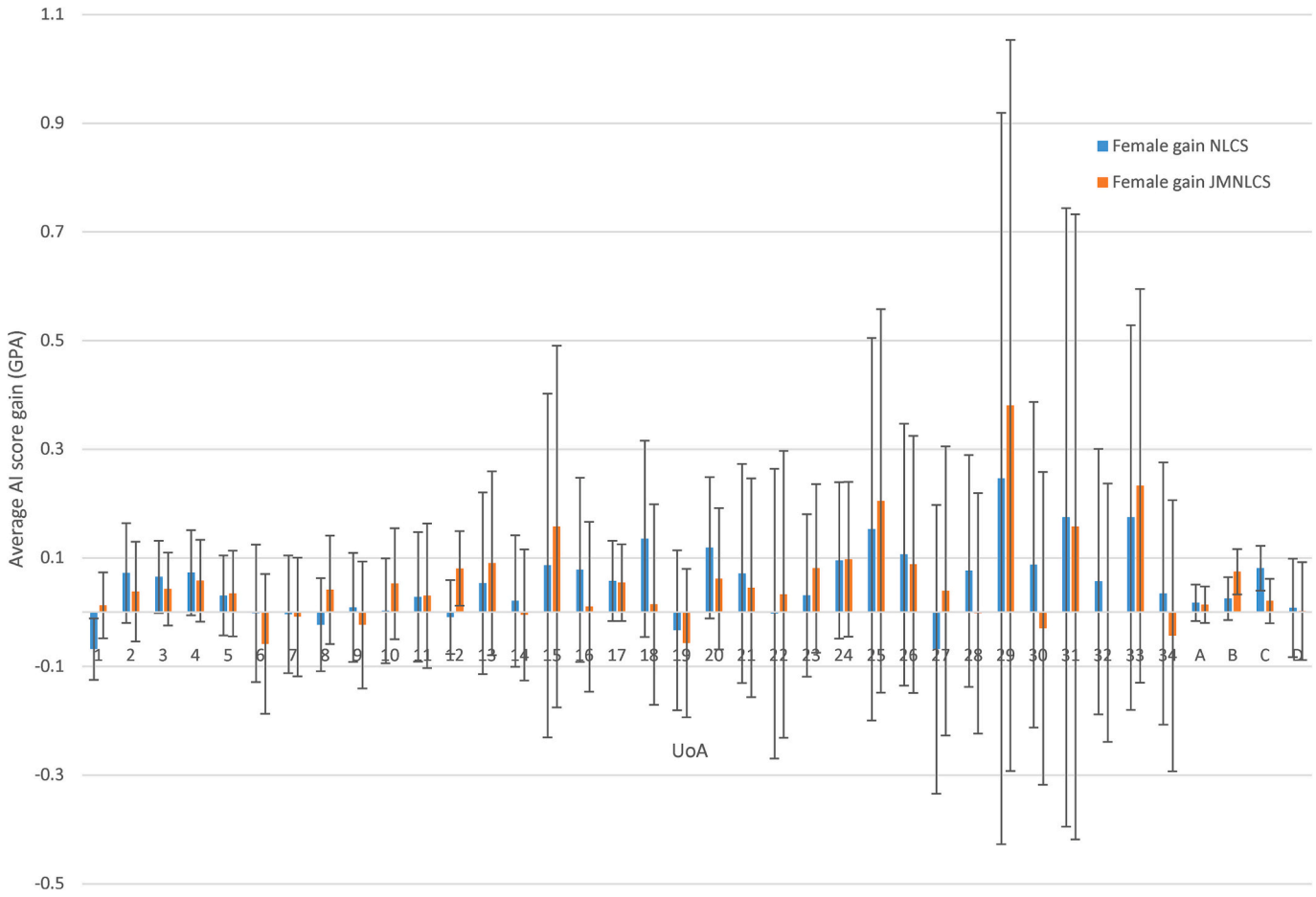

Fig. 2. Prediction gains for female researchers compared to male researchers from allocating scores with NLCS or JMNLCS instead of peer review.

(Fig. 3). Whilst the error bars contain zero in most cases, interdisciplinary research in UoAs 16 (Economics & Econometrics) and 18 (Politics & International Studies) has statistically significant moderate advantages from both NLCS and JMNLCS. There is a weak but statistically significant bibliometric interdisciplinary gain for Main Panel B JMNLCS and Main Panel C NLCS. The overall UoA pattern is not statistically significantly different from equal, however, with 20 out of 34 UoAs having an advantage with NLCS ($p = 0.081$ for a post-hoc binomial test for gender difference $\alpha = 0.5$) and 21 with JMNLCS ($p = 0.054$ for a post-hoc binomial test for gender difference $\alpha = 0.5$).

5. Discussion

The analysis has many limitations. Bibliometrics may have a different value relative to peer review outside the UK and for different peer review goals. UK researchers submitted only their best work, and academics producing outputs thought to be lower quality may have been transferred to teaching contracts to avoid having their outputs assessed (affecting average scores in rankings tables). Other field/year normalised indicators may have produced different results, particularly if they did not take citation skewing into account. The 34 UoAs used for the analysis are relatively broad and a different categorisation scheme may have produced slightly different outcomes. The results may also change over time, and particularly for the journal-level analysis with the continued rise of large broad scope open access megajournals like PLoS One (Spezi et al., 2017). The gender detection may have introduced a second order bias related to ethnicity for names that were not detected with the algorithm. Perhaps most importantly, the interdisciplinary research flag may be inaccurate. It is possible that interdisciplinary differences found are second order effects from large strong or weak HEIs using it differently from average. Longer citation windows are also sometimes needed to assess interdisciplinary research citations (Chen et al., 2022), which may also have been a factor. Finally, the study has not investigated the cause of the gender bias in bibliometrics, relative to peer review, and knowledge of this might help to judge whether the underlying bias is in the peer review or the bibliometrics.

For RQ1, the finding that article-level and journal-level bibliometrics disadvantage high quality departments, compared to peer review, seem to be the first of its kind. Whilst the bibliometric disadvantage for higher scoring departments has a simple and logical explanation (see Results), it does not seem to have been remarked on in previous studies, in evaluation criteria for national research evaluation exercises, or in performance related funding procedures. These results are limited by the bibliometric ties being randomly allocated higher or lower scores, however. Whilst this simulates how the bibliometrics would have to be used if the exact number of articles in each star rating class is predetermined, such a use would be unrealistic in practice unless an assessment had fixed quotas for quality scores, such as to norm reference between fields in the assessment practice. Thus, this random assignment could be the reason why the bibliometrics have a damping effect in most UoAs. Nevertheless, the problem of ties would need to be resolved somehow if bibliometrics were to be used, and there does not seem to be a fairer solution.

For RQ2, the minor gender bias in favour of women from both article-level and journal-level citations aligns with prior research of small gender citation advantage of women compared to men in the UK (Thelwall, 2020). It extends this by suggesting that citations slightly overestimate the quality of female first-authored research, as judged by

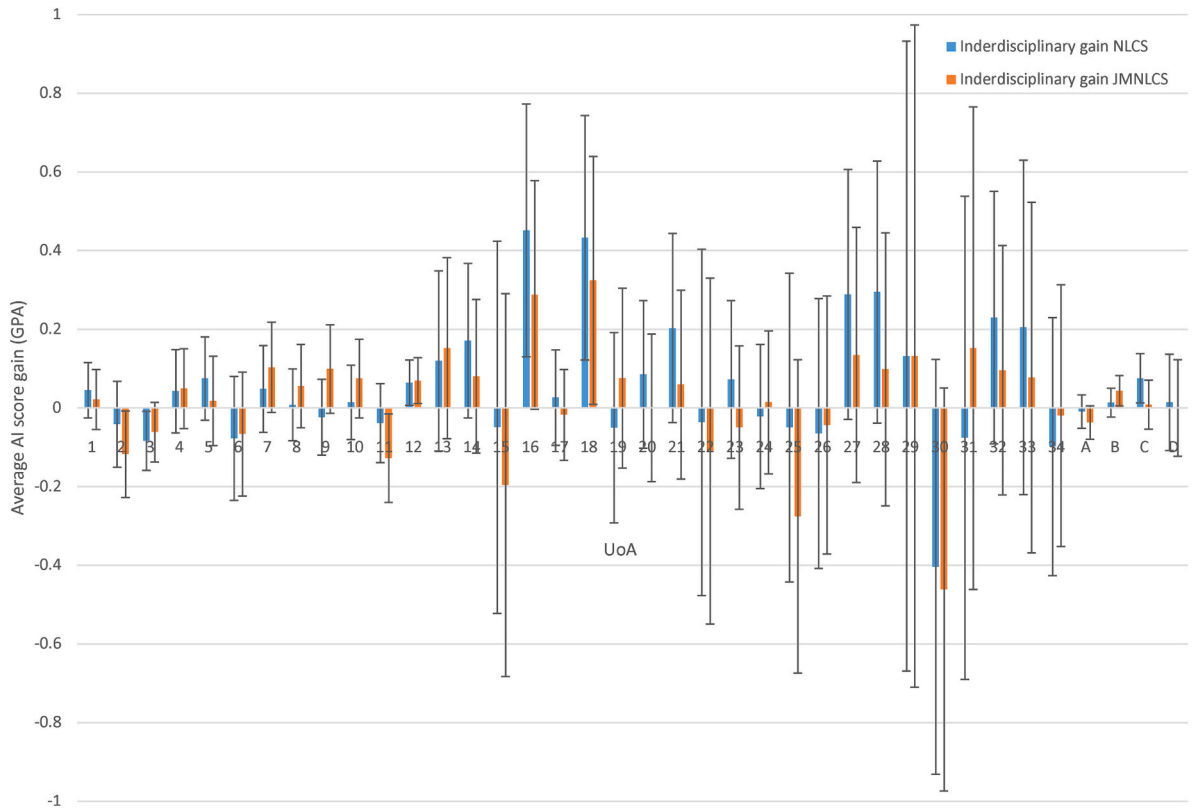

Fig. 3. Score gains for interdisciplinary research compared to monodisciplinary research from allocating scores with NLCS or JMNLCS instead of peer review.

peer review. This partly conflicts with a previous suggestion that citations underestimate the significance of female first authored research because it tends to be more read than cited (Thelwall, 2018). Thus, either the peer review scores have a slight male bias, such as by insufficiently considering wider societal value, or the previous argument based on readership information was incorrect, perhaps because it did not consider non-educational impacts, such as commercial value. The results agree with a related finding for Italy 2004–2010 that used a bibliometric heuristic combining journal impact factors and citation and also factored out age and seniority and rates and used research fields as dummy variables in a combined regression (Jappelli et al., 2017; see also: HEFCE, 2015). It is also possible that the results hide other bibliometric gender biases through gender differences in team contributions. For example, perhaps bibliometrics favour senior male last authors compared to peer review.

For RQ3, the lack of an overall trend in the relationship between interdisciplinarity and any citation advantage is the first result of its kind but aligns with prior arguments that interdisciplinarity is complex, with no simple quality pattern (Huutoniemi, 2010) including for its citation relationship (Yegros-Yegros et al., 2015). The existence of exceptions in relatively small UoAs may be due to relatively stable interdisciplinary fields, such as econophysics, with high levels of citation (Sharma and Khurana, 2021). Recall that the interdisciplinary evidence used in the current article was partial, however, so it is possible that a relationship exists but has been hidden by the method of flagging interdisciplinary research.

6. Conclusion

This study found that departments producing better research (as judged by peer review) tend to be disadvantaged when bibliometrics are used, even in fields where bibliometrics have high correlations with quality scores. This may be due to the damping effect of randomly assigning tied bibliometric scores to higher or lower classes. Thus, evaluation exercises relying on bibliometrics should be aware of this potential deficiency and either accept it or take steps to remedy it. This applies equally to exercises, like the REF, where bibliometrics are used to support peer review rather than to replace it. For example, if the bibliometric information does not help make a quality decision in cases where REF peer reviewers disagree, it would be logical to favour a quality score that aligned with the departmental average. This would give a small nudge to partly offset the bibliometric bias.

The minor gender advantage for females compared to males for bibliometrics in the UK should be reassuring for those seeking to use bibliometrics to support research assessment in the sense that it is unlikely to introduce a bias against women, at least compared to peer review. Given the additional obstacles faced by women in society and academia, a small citation bias in their favour may help to reduce systemic biases against them.

The results also suggest that interdisciplinary research is not disadvantaged overall by bibliometrics compared to peer review. Nevertheless, evaluators should be watchful for individual high or low citation interdisciplinary fields in which bibliometrics may be misleading.

For individual-level research evaluations consulting bibliometrics, such as for appointments, promotions and tenure, the results suggest that article-level and journal-level citation rate information will not disadvantage women or interdisciplinary researchers overall. This supports the continued use of article-level bibliometrics in these contexts, when appropriate.

CRediT authorship contribution statement

Mike Thelwall: Methodology, Analysis, Writing–original draft.
Kayvan Kousha: Writing–review & editing.
Mahshid Abdoli: Writing–review & editing.
Emma Stuart: Writing–review & editing.
Meiko Makita: Writing–review & editing.
Paul Wilson: Writing–review & editing.

Jonathan Levitt: Writing–review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: MIKE THELWALL reports financial support was provided by Research England, Scottish Funding Council, Higher Education Funding Council for Wales, and Department for the Economy, Northern Ireland.

Data availability

The funders required the raw data to be deleted.

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 funders had no role in the design or execution of this study. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

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 Serv. Res. 42 (1p1), 329–346.

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

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.

Ancaiani, A., Anfossi, A.F., Barbara, A., Benedetto, S., Blasi, B., Carletti, V., Sileoni, S., 2015. Evaluating scientific research in Italy: the 2004–10 research evaluation exercise. Res. Eval. 24 (3), 242–255.

Andersen, J.P., Schneider, J.W., Jagsi, R., Nielsen, M.W., 2019. Meta-research: gender variations in citation distributions in medicine are very small and due to self-citation and journal prestige. Elife 8, e45374.

ARC, 2022. New Working Group to advise on ERA Transition. https://www.arc.gov.au/news-publications/media/media-releases/new-working-group-advise-era-transition.

Arnold, A., Cafer, A., Green, J., Haines, S., Mann, G., Rosenthal, M., 2021. Perspective: promoting and fostering multidisciplinary research in universities. Res. Policy 50 (9), 104334.

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

Banda, R.M., 2020. From the inside looking out: Latinas intersectionality and their engineering departments. Int. J. Qual. Stud. Educ. 33 (8), 824–839.

Begeny, C.T., Ryan, M.K., Moss-Racusin, C.A., Ravetz, G., 2020. In some professions, women have become well represented, yet gender bias persists—perpetuated by those who think it is not happening. Science. Advances 6 (26), eaba7814.

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

Beshyah, S.A., 2019. Authors' selection of target journals and their attitudes to emerging journals: a survey from two developing regions. Sultan Qaboos Univ. Med. J. 19 (1), e51.

Biggs, J., Hawley, P.H., Biernat, M., 2018. The academic conference as a chilly climate for women: effects of gender representation on experiences of sexism, coping responses, and career intentions. Sex Roles 78 (5), 394–408.

Boyack, K.W., Klavans, R., 2010. Co-citation analysis, bibliographic coupling, and direct citation: which citation approach represents the research front most accurately? J. Am. Soc. Inf. Sci. Technol. 61 (12), 2389–2404.

Bryce, C., Dowling, M., Lucey, B., 2020. The journal quality perception gap. Res. Policy 49 (5), 103957.

Ceci, S.J., Williams, W.M., 2011. Understanding current causes of women's underrepresentation in science. Proc. Natl. Acad. Sci. 108 (8), 3157–3162.

Chen, S., Song, Y., Shu, F., Larivière, V., 2022. Interdisciplinarity and impact: the effects of the citation time window. Scientometrics 127 (3), 2621–2642.

Currie, R.R., Pandher, G.S., 2020. Finance journal rankings: active scholar assessment revisited. J. Bank. Financ. 111, 105717.

DORA, 2023. San Francisco Declaration on Research Assessment. https://sfdora.org/.

Downes, B.J., Lancaster, J., 2019. Celebrating women conducting research in freshwater ecology… and how the citation game is damaging them. Mar. Freshw. Res. 71 (2), 139–155.

van Driel, M.L., Maier, M., Maeseneer, J.D., 2007. Measuring the impact of family medicine research: scientific citations or societal impact? Fam. Pract. 24 (5), 401–402.

Drivas, K., Kremmydas, D., 2020. The Matthew effect of a journal's ranking. Res. Policy 49 (4), 103951.

Else, H., 2018. Can a major AI conference shed its reputation for hosting sexist behaviour? Nature 563 (7731), 610–612.

EU, 2022. Agreement on reforming research assessment. https://coara.eu/app/uploads/2022/09/2022_07_19_rra_agreement_final.pdf.

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

Fuchs, M.Ž., 2014. Bibliometrics: Use and abuse in the humanities. In: Blockmans, Wim, Engwall, Lars, Weaire, Denis (Eds.), Bibliometrics. Portland Press, Use and Abuse in the Review of Research Performance, Portland, OR, pp. 107–116.

Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., Trow, M., 1994. The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies. Sage, Thousand Oaks, CA.

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

HEFCE, 2015. The Metric Tide: Correlation Analysis of REF2014 Scores and Metrics (Supplementary Report II to the Independent Review of the Role of Metrics in Research Assessment and Management). Higher Education Funding Council for England. https://www.ukri.org/publications/review-of-metrics-in-research-assessment-and-management/.

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

Huang, J., Gates, A.J., Sinatra, R., Barabási, A.L., 2020. Historical comparison of gender inequality in scientific careers across countries and disciplines. Proc. Natl. Acad. Sci. 117 (9), 4609–4616.

Hurtado, J., Pinzón-Fuchs, E., 2021. Understanding the Effects of Journal Impact Factors on the Publishing Behavior of Historians of Economics. Œconomia. History, Methodology, Philosophy, 11-3, pp. 485–496.

Huutoniemi, K., 2010. Evaluating interdisciplinary research. In: Frodeman, K.M. (Ed.), The Oxford Handbook of Interdisciplinarity, vol. 10. Oxford University Press, Oxford, pp. 309–320.

Jain, A., Khor, K.S., Beard, D., Smith, T.O., Hing, C.B., 2021. Do journals raise their impact factor or SCImago ranking by self-citing in editorials? A bibliometric analysis of trauma and orthopaedic journals. ANZ J. Surg. 91 (5), 975–979.

Jappelli, T., Nappi, C.A., Torrini, R., 2017. Gender effects in research evaluation. Res. Policy 46 (5), 911–924.

Jisc, 2022. Evaluating research assessment. https://www.jisc.ac.uk/future-research-assessment-programme/evaluation-activities.

Jump, P., 2015. Can the Research Excellence Framework run on metrics? Times Higher Education. https://www.timeshighereducation.com/can-the-research-excellence-framework-ref-run-on-metrics.

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

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

Larivière, V., Desrochers, N., Macaluso, B., Mongeon, P., Paul-Hus, A., Sugimoto, C.R., 2016. Contributorship and division of labor in knowledge production. Soc. Stud. Sci. 46 (3), 417–435.

Lee, C.J., Sugimoto, C.R., Zhang, G., Cronin, B., 2013. Bias in peer review. J. Am. Soc. Inf. Sci. Technol. 64 (1), 2–17.

Lei, L., Sun, Y., 2020. Should highly cited items be excluded in impact factor calculation? The effect of review articles on journal impact factor. Scientometrics 122 (3), 1697–1706.

Levitt, J.M., Thelwall, M., 2008. Is multidisciplinary research more highly cited? A macrolevel study. J. Am. Soc. Inf. Sci. Technol. 59 (12), 1973–1984.

Lin, C.S., 2018. An analysis of citation functions in the humanities and social sciences research from the perspective of problematic citation analysis assumptions. Scientometrics 116 (2), 797–813.

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

Mackinnon, S., Drozdowska, B.A., Hamilton, M., Noel-Storr, A.H., McShane, R., Quinn, T., 2018. Are methodological quality and completeness of reporting associated with citation-based measures of publication impact? A secondary analysis of a systematic review of dementia biomarker studies. BMJ Open 8 (3), e020331.

Mahdi, S., D'Este, P., Neely, A., 2008. Citation Counts: Are they Good Predictors of RAE Scores? Advanced Institute of Management Research, London. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1154053.

Maliniak, D., Powers, R., Walter, B.F., 2013. The gender citation gap in international relations. Int. Organ. 67 (4), 889–922.

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

Meese, K.A., O'Connor, S.J., Borkowski, N., Hernandez, S.R., 2017. Journal rankings and directions for future research in health care management: a global perspective. Health Serv. Manag. Res. 30 (2), 129–137.

Meho, L.I., 2021. The gender gap in highly prestigious international research awards, 2001–2020. Quant. Sci. Stud. 2 (3), 976–989.

Merton, R.K., 1968. The Matthew effect in science: the reward and communication systems of science are considered. Science 159 (3810), 56–63.
Merton, R.K., 1973. The Sociology of Science: Theoretical and Empirical Investigations. University of Chicago press, Chicago, IL.
Mingers, J., Yang, L., 2017. Evaluating journal quality: a review of journal citation indicators and ranking in business and management. Eur. J. Oper. Res. 257 (1), 323–337.
Mryglod, O., Kenna, R., Holovatch, Y., Berche, B., 2015. Predicting results of the research excellence framework using departmental h-index. Scientometrics 102 (3), 2165–2180.
Nieminen, P., Carpenter, J., Rucker, G., Schumacher, M., 2006. The relationship between quality of research and citation frequency. BMC Med. Res. Methodol. 6 (1), 1–8.
Overholtzer, L., Jalbert, C.L., 2021. A "leaky" pipeline and chilly climate in archaeology in Canada. Am. Antiq. 86 (2), 261–282.
Pölönen, J., Guns, R., Kulczycki, E., Sivertsen, G., Engels, T.C., 2021. National lists of scholarly publication channels: an overview and recommendations for their construction and maintenance. J. Data Inform. Sci. 6 (1), 50–86.
Pride, D., Knoth, P., 2018. Peer review and citation data in predicting university rankings, a large-scale analysis. In: International Conference on Theory and Practice of Digital Libraries. Springer, Berlin, Germany, pp. 195–207.
van Raan, A.F., 2004. Measuring science. In: Handbook of Quantitative Science and Technology Research. Springer, Dordrecht, pp. 19–50.
REF, 2019. Interdisciplinary Research. https://www.ref.ac.uk/about-the-ref/interdisciplinary-research/.
REF, 2020. Panel criteria and working methods. https://www.ref.ac.uk/media/1450/ref-2019_02-panel-criteria-and-working-methods.pdf.
Rinia, E.J., Van Leeuwen, T.N., Van Vuren, H.G., Van Raan, A.F., 1998. Comparative analysis of a set of bibliometric indicators and central peer review criteria: evaluation of condensed matter physics in the Netherlands. Res. Policy 27 (1), 95–107.
Rubin, M., 2017. Do p values lose their meaning in exploratory analyses? It depends how you define the familywise error rate. Rev. Gen. Psychol. 21 (3), 269–275.
Seglen, P.O., 1997. Why the impact factor of journals should not be used for evaluating research. Bmj 314 (7079), 497.
Serenko, A., Bontis, N., 2021. Global ranking of knowledge management and intellectual capital academic journals: a 2021 update. J. Knowl. Manag. 26 (1), 126–145.
Sharma, K., Khurana, P., 2021. Growth and dynamics of Econophysics: a bibliometric and network analysis. Scientometrics 126 (5), 4417–4436.
Sivertsen, G., 2017. Unique, but still best practice? The research excellence framework (REF) from an international perspective. Palgrave Commun. 3 (1), 1–6.
Sivertsen, G., 2023. Performance-based research funding and its impacts on research organizations. In: Handbook of Public Funding of Research. Edward Elgar Publishing, pp. 90–106.
de Solla Price, D., 1976. A general theory of bibliometric and other cumulative advantage processes. J. Am. Soc. Inf. Sci. 27 (5), 292–306.
Sønderstrup-Andersen, E., Sønderstrup-Andersen, H., 2008. An investigation into diabetes researcher's perceptions of the journal impact factor—reconsidering evaluating research. Scientometrics 76 (2), 391–406.
Spezi, V., Wakeling, S., Pinfield, S., Creaser, C., Fry, J., Willett, P., 2017. Open-access mega-journals: the future of scholarly communication or academic dumping ground? A review. J. Doc. 73 (2), 263–283. https://doi.org/10.1108/JD-06-2016-0082.
Stirling, A., 2007. A general framework for analysing diversity in science, technology and society. J. R. Soc. Interface 4 (15), 707–719.
Stoet, G., Geary, D.C., 2018. The gender-equality paradox in science, technology, engineering, and mathematics education. Psychol. Sci. 29 (4), 581–593.
Stokes, D.E., 1997. Pasteur's Quadrant: Basic Science and Technological Innovation. Brookings Institution Press, New York, NY.
Sugimoto, C.R., Weingart, S., 2015. The kaleidoscope of disciplinarity. J. Doc. 71 (4), 775–794.
Thelwall, M., 2017. Three practical field normalised alternative indicator formulae for research evaluation. J. Informetrics 11 (1), 128–151. https://doi.org/10.1016/j.joi.2016.12.002.
Thelwall, M., 2018. Do females create higher impact research? Scopus citations and Mendeley readers for articles from five countries. J. Informetrics 12 (4), 1031–1041.
Thelwall, M., 2020. Female citation impact superiority 1996-2018 in six out of seven English-speaking nations. J. Assoc. Inf. Sci. Technol. 71 (8), 979–990. https://doi.org/10.1002/asi.24316.
Thelwall, M., Fairclough, R., 2015. Geometric journal impact factors correcting for individual highly cited articles. J. Informetrics 9 (2), 263–272.
Thelwall, M., Mas-Bleda, A., 2020. A gender equality paradox in academic publishing: countries with a higher proportion of female first-authored journal articles have larger first author gender disparities between fields. Quant. Sci. Stud. 1 (3), 1260–1282.
Thelwall, M., Bailey, C., Tobin, C., Bradshaw, N., 2019. Gender differences in research areas, methods and topics: can people and thing orientations explain the results? J. Informetrics 13 (1), 149–169.
Thelwall, M., Abdoli, M., Lebiedziewicz, A., Bailey, C., 2020. Gender disparities in UK research publishing: differences between fields, methods and topics. Profesional de la Información 29 (4), e290415. https://doi.org/10.3145/epi.2020.jul.15.
Traag, V.A., Waltman, L., 2019. Systematic analysis of agreement between metrics and peer review in the UK REF. Palgrave Commun. 5 (1), 29.
UNESCO, 2022. Gender equality: How global universities are performing. https://www.iesalc.unesco.org/wp-content/uploads/2022/03/SDG5_Gender_Report-2.pdf.
Van Raan, A.F., 2006. Statistical properties of bibliometric indicators: Research group indicator distributions and correlations. J. Am. Soc. Inf. Sci. Technol. 57 (3), 408–430.
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. J. Informetrics 5 (1), 14–26.
Walters, W.H., 2017. Do subjective journal ratings represent whole journals or typical articles? Unweighted or weighted citation impact? J. Informetrics 11 (3), 730–744.
Waltman, L., Traag, V., 2020. Use of the Journal Impact Factor for Assessing Individual Articles: Statistically Flawed or Not? F1000Research, 9.
Wang, J., 2013. Citation time window choice for research impact evaluation. Scientometrics 94 (3), 851–872.
Wang, J., Veugelers, R., Stephan, P., 2017. Bias against novelty in science: a cautionary tale for users of bibliometric indicators. Res. Policy 46 (8), 1416–1436.
Wang, X., Dworkin, J.D., Zhou, D., Stiso, J., Falk, E.B., Bassett, D.S., Lydon-Staley, D.M., 2021. Gendered citation practices in the field of communication. Ann. Int. Commun. Assoc. 45 (2), 134–153.
West, R., McIlwaine, A., 2002. What do citation counts count for in the field of addiction? An empirical evaluation of citation counts and their link with peer ratings of quality. Addiction 97 (5), 501–504.
Whitley, R., 2000. The Intellectual and Social Organization of the Sciences. Oxford University Press, Oxford, UK.
Wilkins-Yel, K.G., Hyman, J., Zounlome, N.O., 2019. Linking intersectional invisibility and hypervisibility to experiences of microaggressions among graduate women of color in STEM. J. Vocat. Behav. 113, 51–61.
Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., 2015. The metric tide. https://www.ukri.org/publications/review-of-metrics-in-research-assessment-and-management/.
Yegros-Yegros, A., Rafols, I., D'este, P., 2015. Does interdisciplinary research lead to higher citation impact? The different effect of proximal and distal interdisciplinarity. PLoS One 10 (8), e0135095.
Zhang, L., Sivertsen, G., Du, H., Huang, Y., Glänzel, W., 2021. Gender differences in the aims and impacts of research. Scientometrics 126 (11), 8861–8886.