Journal metrics as predictors of Research Excellence Framework 2021 results: Comparison of impact factor quartiles and Finnish expert-ratings

Janne Pölönen[*], Raf Guns[**] and Tim Engels[**]

[*]*janne.polonen@tsv.fi*
https://orcid.org/0000-0003-1649-0879
Publication Forum, Federation of Finnish Learned Societies, Finland

[**] *raf.guns@uantwerpen.be; tim.engels@uantwerpen.be*
https://orcid.org/0000-0003-3129-0330; https://orcid.org/0000-0002-4869-7949
Centre for R&D Monitoring, University of Antwerp, Belgium

This study compares citation-based and expert-based journal metrics as predictors of peer-assessed research quality based on 154,826 journal articles submitted to UK's Research Excellence Framework (REF) 2021. The Finnish expert-based Julkaisufoorumi (JUFO) level ratings of journals determined by expert-panels per field produce scores that correlate more strongly with REF scores than those based on citation-based Journal Impact Factor (JIF) or Journal Citation Indicator (JCI) Quartiles. This holds true at aggregate levels of 34 Subject areas, 157 Higher Education Institutions (HEI), and 1,888 Units of Assessment (UoA). Especially non-field-normalised JIF-based scores correlate poorly with REF scores. All types of journal metrics are more aligned with expert-based REF scores at the highest aggregate level of HEIs and agree less at the lower aggregate level of UoAs and Subject areas.

1. Introduction

This study compares citation-based and expert-based journal metrics as predictors of peer-assessed research quality based on 154,826 journal articles submitted to UK's Research Excellence Framework (REF) 2021. REF2021 is the latest campaign of the performance-based research funding system established in 1992. Like the preceding iterations, REF2021 relied exclusively on evaluation of individual outputs by field-specific expert panels as the golden standard for determining research quality (Wilsdon et al., 2015).

Several studies have used the REF results to investigate correlations between peer review and metrics in research assessment (Thelwall et al., 2022; Waltman & Traag, 2021). Our aim is to investigate the relation of three journal metrics to research quality as determined by REF expert panels:

- Journal Impact Factor (JIF) Quartiles are determined per WoS subject category. We focus on JIF because "despite the well-known technical and interpretive concerns, the JIF remains the standard journal indicator" (Larivière & Sugimoto, 2019). Moreover, several countries use JIF quartiles in research assessment, despite their drawbacks (Vîiu & Păunescu, 2021).
- Journal Citation Indicator (JCI) Quartiles are a field-normalized journal-level indicator", introduced in the Journal Citation Reports 2021 (Szomszor, 2021; Torres-Salinas et al., 2022).
- The Julkaisufoorumi (JUFO) classification has four levels (1 = basic, 2 = leading, 3 = top, 0 = other) for domestic and foreign peer-reviewed publication channels (journals and book publishers) determined by Finnish experts in 23 field-specific panels (Pölönen & Auranen, 2022).

We apply JIF Quartiles, JCI Quartiles and JUFO levels to 154,826 journal articles submitted to REF2021 to analyse their agreement with peer review results of REF panels at the aggregate

levels of 34 Subject areas, 157 UK Higher Education Institutions (HEI), and 1,888 Units of Assessment (UoA). Our research questions are:

1. What is the coverage and distribution across JIF quartiles, JCI Quartiles and JUFO-levels of the journal articles submitted to REF2021?
2. To what extent do REF-based scores for universities, UoA and fields agree with scores based on JIF Quartiles, JCI Quartiles and JUFO levels?

Our aim is not to investigate if expert assessment by REF panels could or should be replaced or informed by journal metrics. The more modest purpose of our study is to contribute to a better understanding of the limitations and – consequently – more responsible use of journal-based metrics in research assessment. In many parts of the world assessments have been too narrowly focused on metrics, in particular the JIF (McKiernan et al., 2019). Such assessments create goal-displacing incentives for publishing peer-reviewed journal articles in high JIF journals.

In 2012, the use of JIF "as a surrogate measure of the quality of individual research articles" was explicitly addressed by the San Francisco Declaration on Research Assessment (DORA). At the same time, the Nordic countries have developed alternative and more inclusive journal-based metrics, which cover different publication types and languages, to be used at macro level for allocating research funding to universities (Pölönen et al., 2020). Another approach has focused on improving citation-based journal indicators, which has led to field-normalized indicators like the Mean Normalized Journal Score (MNJS).

Avoidance of inappropriate use of publication- and journal-based metrics is one of the key changes advocated by the new international Agreement on Reforming Research Assessment. "Responsible use of quantitative indicators can", according to the Agreement, "support assessment where meaningful and relevant, which is context dependent" (CoARA, 2022). In the bibliometric literature, metrics are generally considered more appropriate for the macro-level than micro-level (Glänzel, 2011), or meta-institutional (national) rather than institutional (internal) assessment processes (Moed, 2020). Given that the Agreement encompasses assessments at different levels, from individual researchers and teams to subunits and organisations, an important question remains (Sivertsen & Rushworth, 2023): what are the appropriate uses of metrics, and especially of publication- and journal-based metrics?

The use of journal-based metrics also relates to a broader discussion about the relation between journals and the quality of research. On the one hand, it is argued that journal reputation or JIF are poor predictors of the quality and impact of individual papers (Brembs, 2017). Expert-based journal ratings are sometimes questioned as subjective, political and unscientific measures of research quality. On the other hand, a linear relationship has been observed in some STEM fields between the share of world-leading outputs submitted to REF2014 and the share of outputs in top JIF quartile journals (Koya & Chowdhury, 2017). Studies comparing expert-based assessments of individual outputs and journal metrics suggest that "journal ratings are good predictors of article quality" (Bonaccorsi et al., 2015).

A complementary goal of this study is to better understand the role of expert assessment and metrics in the context of journal evaluation. While there is a need to develop next-generation metrics for diverse outputs of research and open science practices, we want to investigate if expert assessment could be used to improve the traditional journal-based metrics. Although machine-learning techniques can, to a certain extent, predict expert-based level ratings for

journals (Saarela et al, 2016; 2020), the Nordic expert-based ratings of journals are relatively independent of JIF (Kulczycki et al, 2022).

2. Data and methods

Our data consists of the openly available REF 2021 Results datasets: 1.) Quality Profiles and 2.) Outputs (https://results2021.ref.ac.uk). We enriched the Outputs dataset with 2021 JIF and JCI quartiles from 3.) Journal Citation Reports (JCR) and 2022 JUFO level from 4.) JUFO-portal (https://www.julkaisufoorumi.fi/en).

1. The Quality Profiles dataset contains assessment results for all 1,880 Units of Assessment (UoA), from 157 UK higher education institutions (HEI) across 34 subject areas, for four kinds of profiles: outputs, impacts, environment, and overall. We only use the results on outputs. For each UoA, we have the share of outputs classified as 4* (world-leading) 3* (internationally excellent), 2* (recognized internationally), 1* (recognized nationally), and unclassified (below the standard of nationally recognized or not meeting the published definition of research). The dataset does not include quality level per publication, only the aggregated shares of quality levels for each UoA.
2. The Outputs dataset contains bibliographic metadata of all 185,353 outputs HEIs submitted to REF2021, including the name of HEI and UoA. In this study we use only journal articles (output type D) because this is the only type of output we can link with both JIF Quartiles and JUFO levels based on ISSNs. Journal articles represent 83.5% of all outputs.
3. JIF and JCI 2021 Quartiles of the journals were retrieved from the JCR and identified for all journal articles in the Outputs dataset based on ISSN. JIF Quartiles were available for 12,202 journals included in Science Citation Index (SCI) and Social Sciences Citation Index (SSCI). The number of journals per quartile was: Q1=3,320, Q2=3,102, Q3=2,916 and Q4=2,864. JCI Quartiles were available for 20,900 journals included in SCI, SSCI, Arts & Humanities Citation Index (AHCI) and Emerging Sources Citation Index (ESCI). JCI Quartiles were distributed as follows: Q1=4,857, Q2=4988, Q3=5,277 and Q4=5,778.
4. JUFO 2022 levels of the journals were retrieved from JUFO-portal and identified for all journal articles in the Outputs dataset based on journals' ISSN codes. JUFO levels were available for 32,724 serials, with the following distribution: 3(Top)=745, 2(Leading)=2,461, 1(Basic)=22,931, 0 (Not fulfilling level 1 criteria)=6,587.

The following steps were taken to prepare the datasets for analyses:

1. Calculate share of REF 4*, 3*, 2*, 1* and Not qualified outputs for each HEI and Subject area, based on UoA data in the Quality Profiles dataset.
2. Calculate share of JIF and JCI Q1, Q2, Q3, Q4 and No JIF/JCI outputs, as well as JUFO level 3, 2, 1, 0 and No JUFO outputs for UoAs, HEIs and Subject areas, based on enriched Outputs dataset.
3. Calculate REF, JIF, JCI and JUFO scores for UoAs, HEIs and Subject areas by dividing the sum of weighted outputs by the number of outputs (Table 1). There are 21 UoAs that have exceptionally submitted to two different panels, which may have evaluated them differently. When results are aggregated at the level of UoAs, HEIs and Subject areas, 1341 outputs are counted twice. Hence the total number of outputs in Table 1 is 156,167 instead of 154,826.

Table 1. Number of articles and weights used for scoring REF2021 Units of Assessment (UoA), Higher Education Institutions (HEI) and Subject areas based on REF Quality levels and journal metrics JIF, JCI and JUFO.

REF Quality Levels	Number of articles	Weights for scoring
4*	55,812	4
3*	74,822	3
2*	22,712	2
1*	2,261	1
u/c	530	1
All	156,137	

JUFO Levels	Number of articles	Weights for scoring
3	54,638	4
2	50,518	3
1	45,892	2
0	1,029	1
No level	4,090	1
All	156,167	

JIF Quartiles	Number of articles	Weights for scoring
Q1	93,675	4
Q2	30,264	3
Q3	11,920	2
Q4	4,134	1
No JIF Q	16,174	1
All	156,167	

JCI Quartiles	Number of articles	Weights for scoring
Q1	112,558	4
Q2	23,522	3
Q3	7,282	2
Q4	1,367	1
No JCI Q	11,438	1
All	156,167	

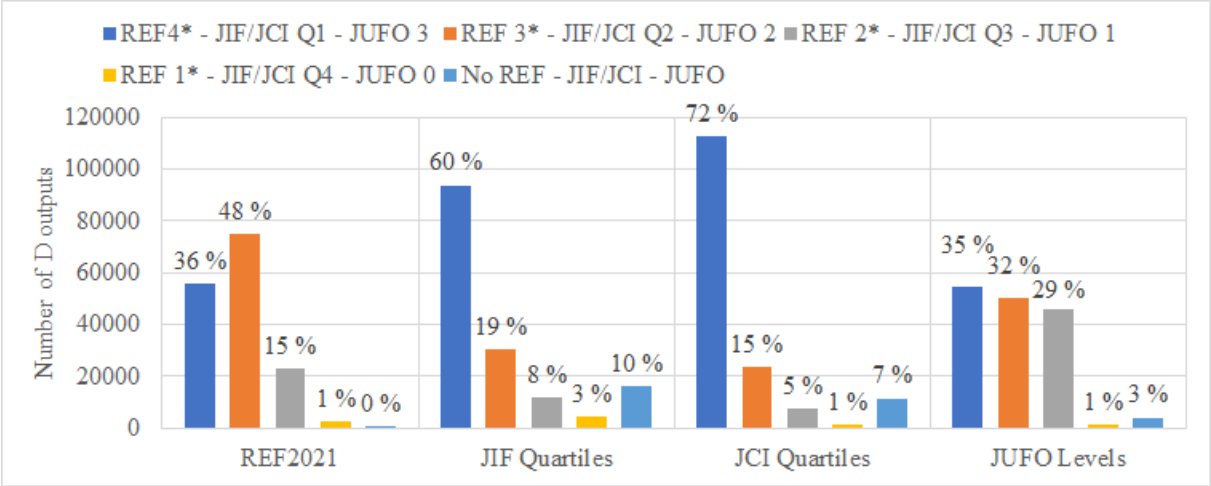
The following analyses are carried out:

1. Distribution of the number of journal articles according to REF quality levels, JIF and JCI Quartiles and JUFO levels.
2. Relation between each UoA's share of REF level 4* outputs and share of JIF/JCI Q1 and Q2, as well as JUFO level 2 and 3, outputs.
3. Correlation between REF scores and JIF/JCI scores, as well as REF scores and JUFO scores, for UoAs, HEIs and Subject-areas. We exclude from this analysis UoAs, HEIs and Subject-areas with less than 20 journal articles and/or less than 50% share of journal articles of all submitted outputs.

3. Results

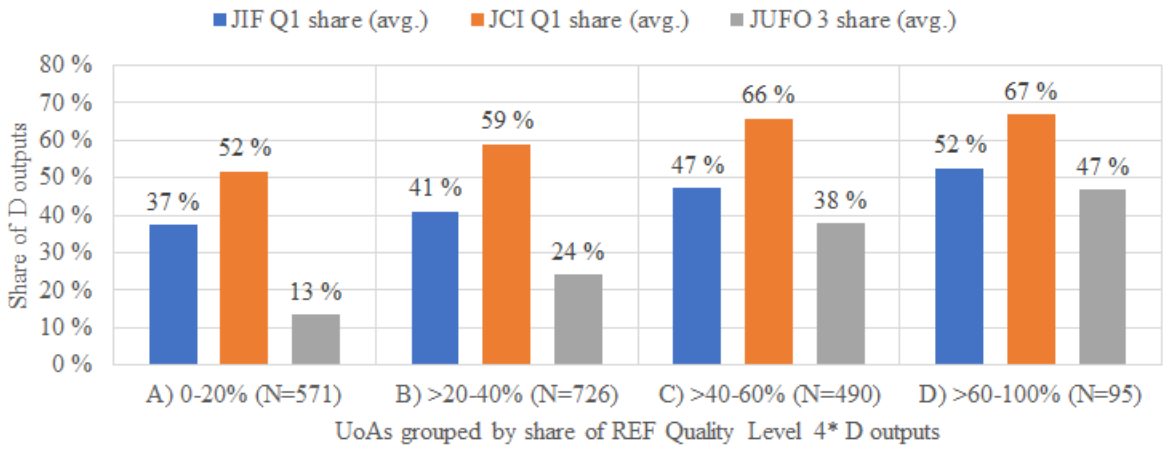
The distribution of journal articles submitted to REF2021 is far more skewed for JIF/JCI Quartiles than for JUFO levels and REF quality levels (Table 1). REF quality levels cover practically all (99.7%), while JUFO levels and JCI Quartiles cover the vast majority of the journal articles (97.4% and 92.7% respectively). JIF Quartiles cover a somewhat smaller share (89.6%).

Figure 1: Distribution of REF2021 journal articles between REF quality levels, JIF/JCI quartiles and JUFO levels.



JIF Q1 and JCI Q1 include the large majority of the journal articles (60% and 72% respectively), while relatively small shares are distributed to lower quartiles (Figure 1). The distribution over JUFO levels 3 to 1, however, is fairly even. This disparity is also due to the fact that JIF or JCI quartiles contain a similar number of journals, whereas the distribution of journals over JUFO levels is very skewed.

Figure 2: UoA share of REF level 4* journal articles compared to average share of JIF/JCI Q1&2 and JUFO level 2&3 journal articles



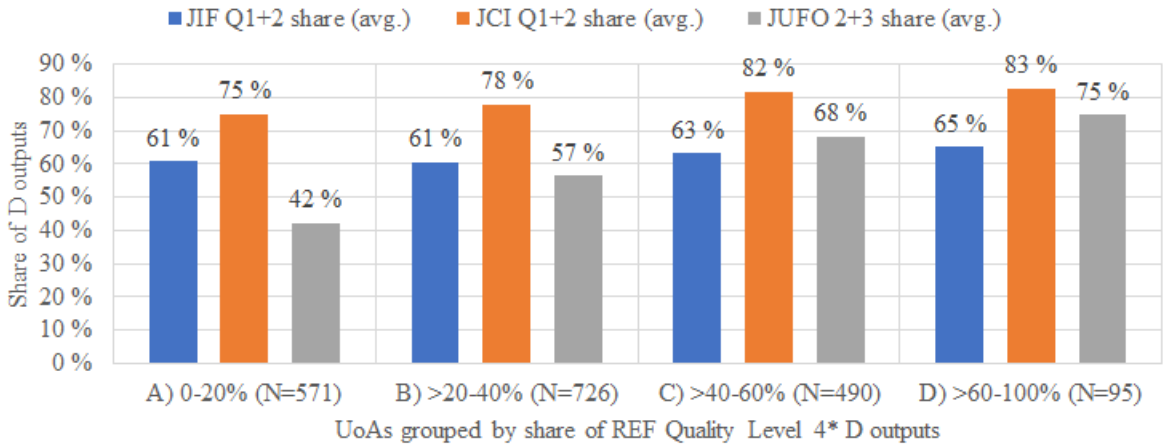


Figure 2 shows a positive relation between the UoAs' share of REF level 4* ("world-leading") journal articles and the JIF and JCI Quartiles and JUFO levels. We distinguish between four groups of UoAs, based on their share of 4* articles, ranging from A (lowest) to D (highest).

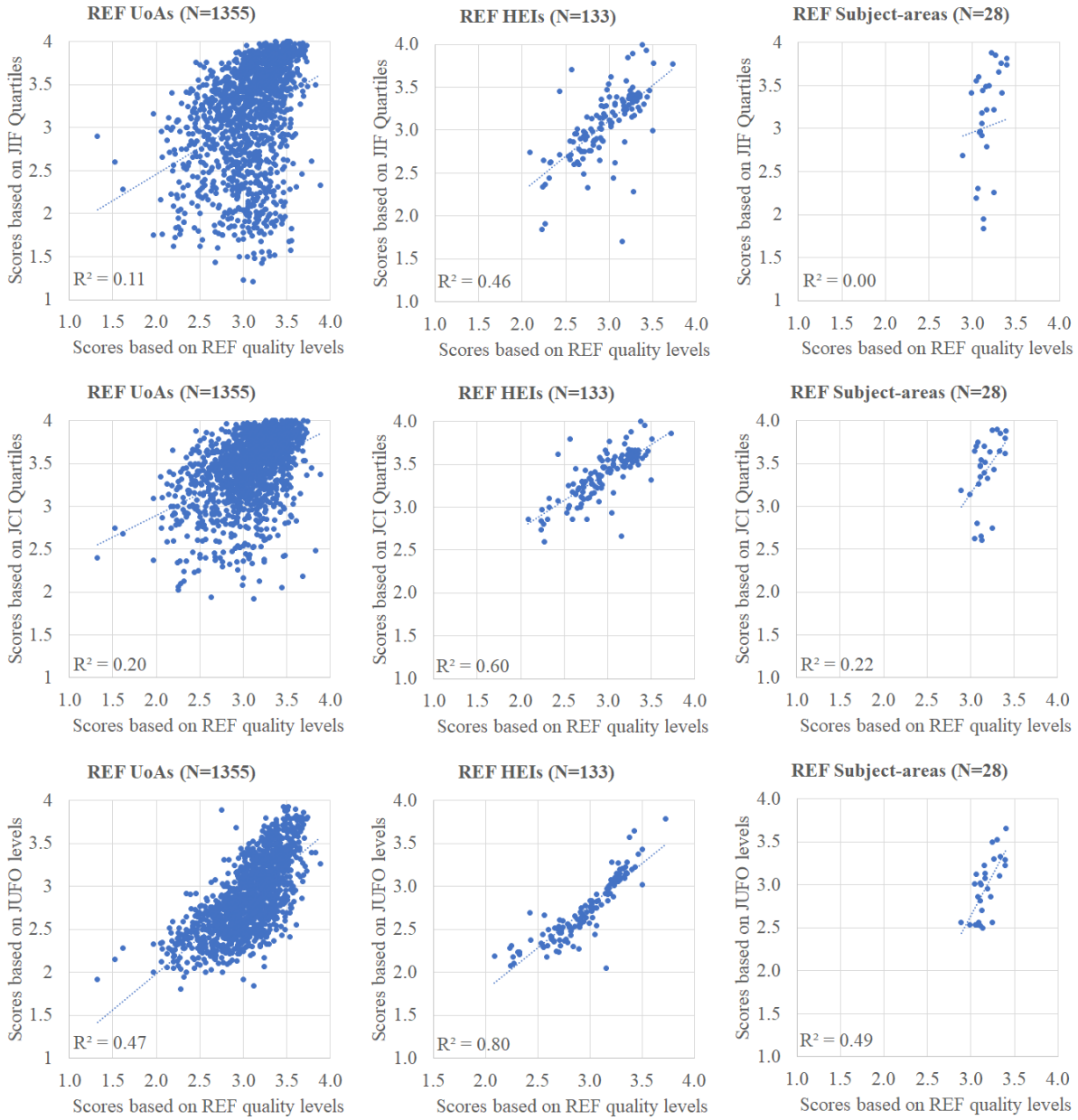
- In the case of JIF, 571 UoAs in group A with less than 20% of 4* journal articles have on average 37% of articles in JIF Q1 journals and 61% in Q1+2 journals; for 95 UoAs in group D with 60% or more 4* outputs the shares are 52% and 65% respectively. The difference between the lowest performing groups A and B is small using JIF Q1.
- JCI differs from JIF mainly in terms of its broader coverage of journals: UoAs in category A have on average 52% of articles in JCI Q1 journals and 75% in Q1+2 journals, whereas in group D the shares are 67% and 83% respectively. The difference between the strongest performing groups C and D is very small using JCI Q1.
- Compared to JIF and JCI Quartiles, the differences between groups A-D in the share of JUFO level 3 and level 2+3 articles appear to be larger. For group A UoAs, the average share of articles in JUFO level 3 journals is 13% but for group D it is 47%, and the shares for JUFO level 2+3 outputs are 42% and 75% respectively.

The shares of JIF and JCI Q1+Q2 do not distinguish well between the four groups.

Correlations between REF scores and scores based on JIF, JCI and JUFO are much stronger at the level of 133 HEIs than at the lower aggregate level of 1,355 UoAs (Figure 3). Correlations at the field level of 28 Subject areas are also low.

- JIF-based scores, which use subject category-based quartiles for journals, show the lowest correlations with REF-based scores at all aggregate levels.
- JCI-based scores, which include humanities as well as emerging journals and are field-normalised, show a stronger correlation with the REF-based scores, especially in the case of HEIs (R^2=0.60).
- JUFO-based scores, which use level ratings of journals by field-specific expert panels, show a much stronger correlation with REF-based scores than the JIF/JCI-based scores, especially at the level of HEIs (R^2=0.80).
- In general, all journal metrics including JUFO, JCI and JIF show a considerably lower correlation with REF-based scores at the level of UoAs.

Figure 3: Correlations between REF-based and JIF-, JCI and JUFO-based scores for REF2021 UoAs, HEIs, and Subject areas.



4. Discussion and conclusions

Our first research question relates to the feasibility of using journal metrics to analyse REF2021 journal articles data in terms of coverage and distribution. Our analysis shows that the Finnish JUFO levels provide almost complete coverage (97.4%) of the journal articles submitted to REF, while JCI and especially JIF provide a more limited coverage (92.7% and 89.6% respectively). The coverage of JIF, JCI and JUFO is high because the REF submissions represent the strongest subset of all journal articles produced by UoAs across different fields, and because over 99.9% of REF journal articles are in English. Most journal articles submitted to REF were published in journals belonging to JIF/JCI Q1. This might be because this is indeed the UoAs' self-selected subset of strongest papers, or because papers were considered the strongest by UoAs because they were published in high impact factor journals.

Expert-based JUFO levels are better able to differentiate UoA with different shares of 4* articles than citation-based JIF and JCI quartiles. This may be a result of JUFO classification's structure, with level 3 being relatively narrow compared to JIF and JCI Q1. Another possible explanation is that expert panels are better able to identify outlets with the most robust editorial practices.

Our second research question was about agreement between REF-based scores and scores based on JIF/JCI Quartiles and JUFO levels. Results show that expert-based level ratings of journals produce results that correlate more strongly with REF scores than those based on JIF/JCI Quartiles across all aggregate levels, although correlations are highest in each case at the level of HEIs.

A limitation of our analysis is that we have not considered differences between fields. In a follow-up study we will investigate the agreement between REF scores and JIF, JCI and JUFO scores, taking fields into account. Nevertheless, Figure 3 shows that JIF-based scores – even if balanced across subject categories – produce large differences between subject areas, whereas their REF scores are relatively similar. The differences between Subject areas are smaller in case of JCI and JUFO scores.

If HEIs were assessed and scored in REF only based on journal articles – which is not the case –, journal-based JUFO levels would produce a ranking order similar to the REF's expert-based assessment of outputs. This is indeed the aggregate level at which JUFO levels are used in Finland: to distribute 14% of core-funding between universities.

Our results show that JUFO-based journal metrics (let alone those based on JCI and JIF) are only partially aligned with expert-based REF assessment at the more granular level of UoAs. It wasn't possible to establish correlations at the level of individual outputs or researchers, but we would expect an even weaker correlation between JUFO scores (or JIF/JCI-based scores) and REF scores. Hence the use of journal-based metrics should be avoided especially at low levels of aggregation.

Finally, it is interesting to consider the possible advantages of expert judgement compared to metrics for journal evaluation. JIFs reduce journal quality to article and citation counts. Journals should be assessed based on multidimensional information regarding, e.g., the integrity and transparency of editorial and peer-review practices (Wouters et al., 2019). Moreover, experts have experience of robustness of editorial practices as authors, reviewers, and editors. As active researchers, they read and use research published in a wide variety of journals, and learn about journals' practices from discussions with colleagues.

Contrary to indicators, expert assessment of journals can "provide a more well-rounded representation of the different dimensions of research quality" (Pölönen et al, 2021). Our results suggest that assessments of journal and article quality by experts in the field may reflect similar dimensions of research quality. Yet journal-based metrics, even if based on expert-assessment, are not suitable for evaluation and comparison of individual researchers. National or institutional level incentives tend to trickle down to individual level, so it is relevant to carefully consider the use of journal-based metrics at higher aggregate levels of assessment.

Open science practices

We reused open datasets from REF2021 and openly available expert-based level ratings for journals from JUFO-portal. Information on JIF and JCI quartiles, however, cannot be shared

openly. We therefore make a dataset openly available that aggregates data at the level of UoA, institution, and subject (Pölönen & Guns, 2023). The figures can be reproduced based on this aggregated dataset, but it allows limited additional analysis.

Author contributions

Janne Pölönen (janne.polonen@tsv.fi): Conceptualization, Methodology, Investigation, Writing original draft, Writing review & editing, Supervision, Project administration. Raf Guns (raf.guns@uantwerpen.be): Conceptualization, Methodology, investigation, Writing review & editing. Tim Engels (tim.engels@uantwerpen.be): Conceptualization, Methodology, Investigation, Writing—review & editing.

Competing interests

Authors have no competing interests.

References

Bonaccorsi, A., Cicero, T., Ferrara, A. & Malgarini, M. (2015). Journal ratings as predictors of articles quality in Arts, Humanities and Social Sciences: an analysis based on the Italian Research Evaluation Exercise. F1000Research 2015, 4:196. https://doi.org/10.12688/f1000research.6478.1

Brembs, B. (2018). Prestigious Science Journals Struggle to Reach Even Average Reliability. Frontiers in Human Neuroscience. https://doi.org/10.3389/fnhum.2018.00037

CoARA (2022). The Agreement full text. https://coara.eu/agreement/the-agreement-full-text/

Glänzel, W. (2011). Thoughts and facts on bibliometric indicators in the light of new challenges in their applications: https://psicologia.ucm.es/data/cont/media/www/807/02%20Glaenzel.pdf

Koya, K. & Chowdhury, G. (2017). Metric-based vs peer-reviewed evaluation of a research output: Lesson learnt from UK's national research assessment exercise. PLoS ONE 12(7): e0179722. https://doi.org/10.1371/journal.pone.0179722

Kulczycki, E., Huang, Y., Zuccala, A., Engels, T., Ferrara, A., Guns, R., Pölönen, J., Sivertsen, G., Taşkın, Z. & Zhang, L. (2022). Uses of the Journal Impact Factor in national journal rankings in China and Europe, Journal of the Association for Information Science and Technology, 73(12), 1741– 1754. https://doi.org/10.1002/asi.24706

Larivière, V. & Sugimoto, C. R. (2019). The Journal Impact Factor: A Brief History, Critique, and Discussion of Adverse Effects. In: Glänzel, W., Moed, H.F., Schmoch, U., Thelwall, M. (eds) Springer Handbook of Science and Technology Indicators. Springer Handbooks. Springer, Cham. https://doi.org/10.1007/978-3-030-02511-3_1

McKiernan, E. C., Schimanski, L. A., Muñoz Nieves, C., Matthias, L., Niles, M. T., Alperin, J. P. (2019). Use of the Journal Impact Factor in academic review, promotion, and tenure evaluations. PeerJ Preprints 7:e27638v2 https://doi.org/10.7287/peerj.preprints.27638v2

Moed, H.F. (2020). Appropriate Use of Metrics in Research Assessment of Autonomous Academic Institutions. Scholarly Assessment Reports, 2(1), p.1. DOI: http://doi.org/10.29024/sar.8

Pölönen, J. & Auranen, O. (2022). Research performance and scholarly communication profile of competitive research funding: the case of Academy of Finland. Scientometrics 127, 7415–7433. https://doi.org/10.1007/s11192-022-04385-8

Pölönen, J. & Guns, R. (2023). Journal metrics as predictors of Research Excellence Framework 2021 results: Comparison of impact factor quartiles and Finnish expert-ratings - dataset. *Zenodo*. https://doi.org/10.5281/zenodo.7837430

Pölönen, J., Guns, R., Kulczycki, E., Sivertsen, G., & Engels, T. C. E. (2020). National lists of scholarly publication channels: An overview and recommendations for their construction and maintenance. Journal of Data and Information Science, 6, 50–86. https://doi.org/10.2478/jdis-2021-0004

Saarela, M., Kärkkäinen, T., Lahtonen, T., & Rossi, T. (2016). Expert-based versus citation-based ranking of scholarly and scientific publication channels. Journal of Informetrics, 10(3), 693–718. https://doi.org/10.1016/j.joi.2016.03.004

Saarela, M., & Kärkkäinen, T. (2020). Can we automate expert-based journal rankings? Analysis of the Finnish publication indicator. Journal of Informetrics, 14(2). doi: https://doi.org/10.1016/j.joi.2020.101008

Sivertsen, G. & Rushworth, A. (2023). The new European reform of research assessment. R-QUEST Policy Brief no. 7. https://www.r-quest.no/wp-content/uploads/2023/02/R-QUEST-Policy-Brief-7.pdf

Szomszor, M. (2021). Introducing the Journal Citation Indicator: A new, field-normalized measurement of journal citation impact. Clarivate Blog, May 20, 2021: https://clarivate.com/blog/introducing-the-journal-citation-indicator-a-new-field-normalized-measurement-of-journal-citation-impact/

Thelwall, M., Kousha, K., Abdoli, M., Stuart, E., Makita, M., Wilson, P. & Levitt, J. (2022). Can REF output quality scores be assigned by AI? Experimental evidence. arXiv:2212.08041. https://doi.org/10.48550/arXiv.2212.08041

Torres-Salinas, D., Valderrama-Baca, P., & Arroyo-Machado, W. (2022). Is there a need for a new journal metric? Correlations between JCR Impact Factor metrics and the Journal Citation Indicator—JCI. *Journal of Informetrics*, 16(3), 101315. https://doi.org/10.1016/j.joi.2022.101315

Vîiu, G.-A., & Păunescu, M. (2021). The lack of meaningful boundary differences between journal impact factor quartiles undermines their independent use in research evaluation. *Scientometrics*, 126(2), 1495–1525. https://doi.org/10.1007/s11192-020-03801-1

Waltman, L., & Traag, V. A. (2021). Use of the journal impact factor for assessing individual articles: Statistically flawed or not? *F1000Research*, 9, 366. https://doi.org/10.12688/f1000research.23418.2

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Kain, R., Kerridge, S., Thelwall, M., Tinkler, J., Viney, I., Wouters, P., Hill, J., & Johnson, B. (2015).

The metric tide. Report of the independent review of the role of metrics in research assessment and management. HEFCE. Retrieved from https://doi.org/10.13140/RG.2.1.4929.1363

Wouters, P., Sugimoto, C. R., Larivière, V., McVeigh, M. E., Pulverer, B., de Rijcke, S., & Waltman, L. (2019). Rethinking impact factors: Better ways to judge a journal. Nature, 569(7758), 621–623. https://doi.org/10.1038/d41586-019-01643-3