Peer review uncertainty at the institutional level

Traag, V.A.; Malgarini, M.; Cicero, T.; Sarlo, S.; Waltman, L.

STI 2018 Leiden

23rd International Conference on Science and Technology Indicators
"Science, Technology and Innovation Indicators in Transition"

STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

Peer review uncertainty at the institutional level

V.A. Traag[*1], M. Malgarini[2], T. Cicero[2], S. Sarlo[2] and L. Waltman[1]

[*] *v.a.traag@cwts.leidenuniv.nl*
[1] Centre for Science and Technology Studies (CWTS), Leiden University, the Netherlands

[2] ANVUR, Rome, Italy

Introduction

Performance based research funding systems vary considerably in how they function (Hicks, 2012). However, they have one element in common: the need to evaluate research. Some countries (such as the UK) have opted for research evaluation that is primarily based on peer review, whereas others (such as Sweden) use a metric driven system to evaluate research. In Italy, the research assessment exercise, known as VQR (*Valutazione della Qualità della Ricerca*), uses an informed peer review approach, where review by carefully selected panelists and external peers is supported by bibliometrics (in suitable fields, see Ancaiani et al., 2015 for more details).

A recurrent question in this context is whether peer review and metrics tend to yield similar outcomes, or whether they differ substantially. This question has been repeatedly addressed in the context of the UK REF (Research Excellence Framework, previously Research Assessment Exercise or RAE) system, culminating in a systematic large-scale comparison between peer review and metrics in the *Metric Tide* report (Wilsdon et al., 2015, Supplementary Report II). We believe this report has two crucial shortcomings (Traag & Waltman, 2017): (1) correlations were studied at the publication level in contrast to the aggregate institutional level at which the REF outcomes are published; and (2) the uncertainty of peer review itself (i.e. the extent to which different peer reviewers or different peer review panels do or do not agree with each other) was not considered. In the Italian context, ANVUR, the agency tasked with the implementation of the VQR, collected data on peer review and was able to quantify peer review uncertainty at the publication level. It found that metrics (especially journal metrics) generally correlate better with peer review than two independent reviewers among each other (Alfò, Benedetto, Malgarini, & Sarlo, 2017).

We here study peer review uncertainty at the institutional level. We rely on data collected by ANVUR that was also used by Alfò, Benedetto, Malgarini and Sarlo (2017). We find that peer review agreement is generally higher at the institutional level than at the publication level. Similarly, correlations between peer review and metrics also tend to be higher at the institutional level. We also find that the correlations between especially journal metrics and peer review are on par with correlations among two peer reviewers.

Data and methods

Data was collected by ANVUR in the framework of the most recent Italian evaluation exercise, relating to the period 2011–14; overall, over 114,000 research items were submitted for evaluation. Here we restrict the analysis to fields that made use of metrics, including all the STEM areas and Economics and Statistics; in other social sciences and humanities no indicators were used to complement peer review. Furthermore, we consider only university institutions, hence excluding public research organisations and other research bodies. After making this selection, a reference population of 58,677 publications remains that were evaluated with the support of metrics. We extract a representative sample stratified according to the distribution of publications by scientific field, called GEV (*Gruppi di Esperti della Valutazione*), and we match the selected publications with the CWTS in-house version of the Web of Science (WoS). The final sample includes 4,560 publications, i.e. almost 8% of the reference population, submitted by 78 Italian universities. Note that universities are not always included in all areas. Alfò et al. (2017) already showed that the sample is statistically representative at the level of scientific areas. Post stratification analysis (see Figure 1) shows that sample representativeness is sufficiently maintained also at the institutional level, since the sample in most cases comprises between 6 and 10% of the reference population.

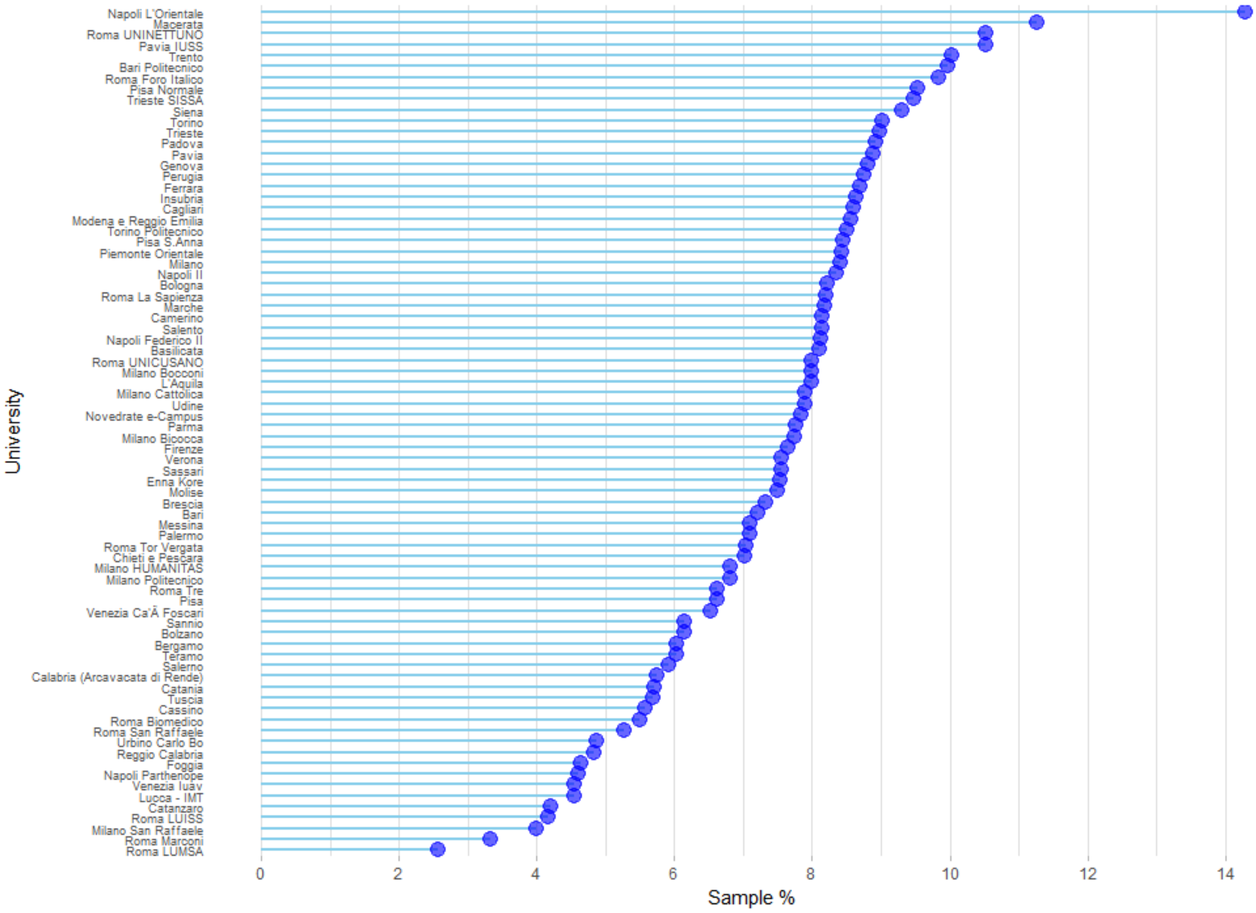

Figure 1 Sample shares at the institutional level.

For each paper included in our sample, we calculate four indicators, two article level metrics and two journal level metrics. First, we calculate the normalised citation score (NCS) for each paper, which is the number of citations divided by the average number of citations of all publications in the same field and in the same year. Secondly, we calculate whether a paper belongs to the top 10% most cited of its field and its year. This indicator is called the P(top 10%). Thirdly, we calculate the normalised journal score (NJS), which is the average NCS of all publications in a certain journal and a certain year. Fourthly, we calculate the proportion of papers in a certain journal and a certain year that belong to the top 10% most cited of their field. This indicator is referred to as the JPP(top 10%). We take into account citations up to (and including) 2015, to be consistent with the timing of the VQR. We use the WoS journal subject categories for calculating normalised indicators. In the case of journals that are assigned to multiple subject categories, we apply a fractionalisation approach to normalise the citations of publications in those journals (Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011). Publications in journals in the multidisciplinary category (e.g. *Science*, *Nature*, *PLOS ONE*) are fractionally reassigned to other subject categories based on their references.

In addition to the specifically collected bibliometric information from the WoS, we also consider the indicators collected by ANVUR during the VQR itself. Those indicators may come from various sources (e.g. Scopus, WoS, MathSciNet), and for different publications different journal indicators may be used (5-year Impact Factor, Article Influence Score, SJR, IPP). All scores are normalised as percentiles with respect to the field definitions as provided by the data source. This procedure allows for a greater degree of flexibility in practice (Anfossi, Ciolfi, Costa, Parisi, & Benedetto, 2016), but also makes the data more heterogeneous, thereby rendering the interpretation of the results more difficult. We nevertheless include these indicators in this study in order to compare them to the bibliometric information obtained exclusively from the WoS.

Each publication is considered by two reviewers. We randomly determine which reviewer is considered reviewer number 1 and which one is considered reviewer number 2. Each reviewer rates a publication on three aspects: (1) originality; (2) rigour; and (3) impact. Each aspect is rated on a scale from 1–10, and the three scores are summed to obtain an overall score that hence ranges from 3–30. We thus obtain for each paper two reviewer scores, the two ANVUR percentile metrics (a citation metric and a journal metric), and the four WoS metrics (two citation metrics and two journal metrics).

We want to compare the correlation between metrics and peer review in a fair way to the internal agreement of peer review. In order to do so, we consistently compare all scores and metrics to the overall score of reviewer 1. Internal peer review agreement is then quantified by the correlation of the overall score of reviewer 2 with the overall score of reviewer 1. Likewise, for each of the metrics, we calculate the correlation between the metric and the overall score of reviewer 1. By performing the analysis in this way, correlations between metrics and peer review can be compared in a fair way to the internal agreement of peer review. If we had chosen to compare each metric to the average score of reviewers 1 and 2, this would have already cancelled out some ‘errors’ in the scores of the reviewers, and as a result the correlations of the metrics with the reviewer scores would not have been directly comparable to the internal peer review agreement.

Finally, at the level of the institutions, we obtain aggregate scores and metrics by taking the average of all scores and metrics at the publication level. There seem to be non-linear correlations between citation metrics and peer reviewer scores, and for that reason we use Spearman correlations throughout this study.

Results

We first provide results at the publication level, see Table 1 and Figures 2 and 3. At first sight, the correlations between metrics and peer review do not seem impressive. They reach 0.41 at most. This correlation is obtained for the NJS indicator for Biology (GEV 5). However, the correlations between reviewer 1 and reviewer 2 are also low, in the order of 0.3, reaching a high of 0.45 for Physics (GEV 2). This shows that it is problematic to judge whether a correlation of 0.3 between peer review and metrics should be considered high or low without comparing this correlation to internal peer review agreement. Overall, the correlations between journal indicators and peer review seem to be slightly higher than or on par with correlations among the two peer reviewers.

The citation metrics generally correlate less well with peer review than the journal metrics. Physics (GEV 2) is an exception to this rule. In Physics, citation metrics generally correlate better with peer review than journal metrics.

Overall, the metrics derived exclusively from the WoS show correlations that are of a similar order of magnitude as the correlations obtained for the more heterogeneous VQR metrics. This shows that there is some robustness in the metrics, and that the origin of the metrics does not seem to be that important.

Nonetheless, we want to emphasise that the correlations depend on the exact way in which indicators are calculated, and that relatively small differences between correlations should not be overinterpreted.

Table 1: Spearman correlations with reviewer 1's score at the publication level

GEV		Number of publications	NCS	P(top 10%)	VQR Cit. Percentile	NJS	JPP(top 10%)	VQR Journal Percentile	Reviewer 2 score
1	Mathematics and Computer Sciences	394	0.28	0.24	0.19	0.37	0.36	0.40	0.36
2	Physics	921	0.44	0.40	0.52	0.36	0.34	0.29	0.45
3	Chemistry	591	0.24	0.11	0.34	0.32	0.31	0.40	0.23
4	Earth Sciences	346	0.19	0.14	0.22	0.40	0.40	0.38	0.29
5	Biology	841	0.30	0.23	0.30	0.41	0.39	0.39	0.27
6	Medicine	1128	0.32	0.27	0.34	0.39	0.39	0.39	0.25
7	Agricultural and veterinary sciences	536	0.28	0.19	0.29	0.32	0.32	0.38	0.33
8b	Civil Engineering	202	0.20	0.03	0.14	0.12	0.11	0.25	0.05
9	Industrial and Information Engineering	726	0.22	0.14	0.25	0.17	0.17	0.22	0.20
11b	Psychology	139	0.26	0.24	0.28	0.40	0.36	0.29	0.25
13	Economics and Statistics	265	0.24	0.17		0.22	0.20		0.32

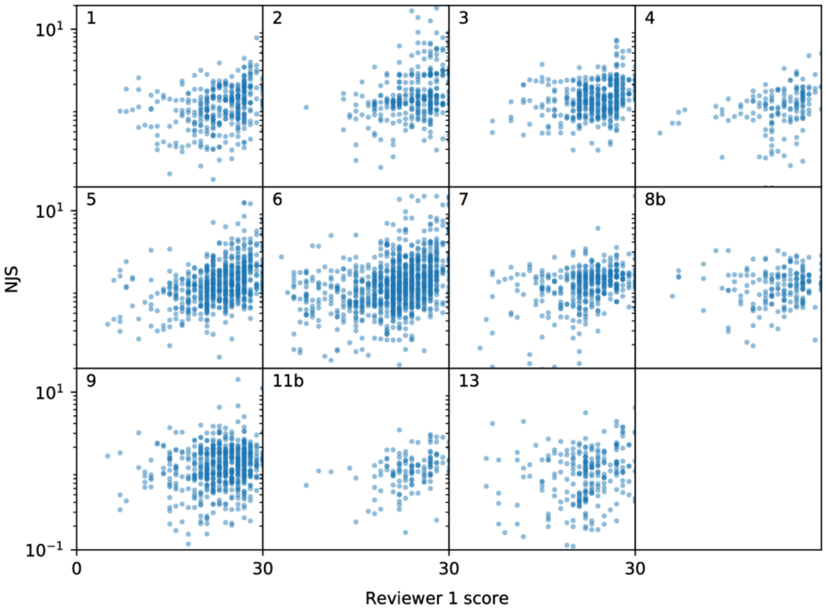

Figure 2 Scatter plot between reviewer 1's score and normalised journal score at the publication level.

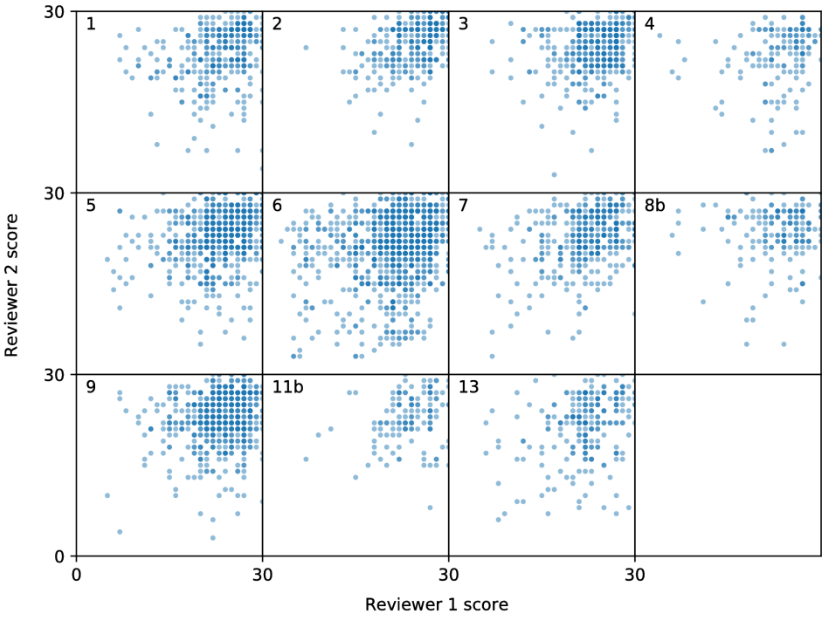

Figure 3 Scatter plot between the scores of reviewer 1 and reviewer 2 at the publication level.

The correlations at the institutional level are generally higher than the correlations at the publication level, see Table 2 and Figures 4 and 5. This suggests that to some extent disagreement between peer review and metrics at the publication level cancels out when moving to the institutional level. In other words, for some publications of an institution, peer review provides a more favourable assessment than metrics, but for other publications, this is the other way around, and therefore peer review and metrics are in stronger agreement at the institutional level than at the publication level. Something similar holds for the agreement among peer reviewers: disagreements among reviewers partly cancel out at the institutional level, and therefore reviewers agree more strongly at this level than at the publication level.

At the publication level discussed above, we found that correlations between journal metrics and peer review were somewhat higher than the correlation between two peer reviewers. This changes at the institutional level, where this remains the case only for Chemistry (GEV 3), Biology (GEV 5) and Civil Engineering (GEV 8b). However, for Civil Engineering, the agreement among the two peer reviewers is extremely low, and the correlation with journal metrics remains as low as about 0.2. For Biology, the correlation of peer review with NJS (0.45) is only slightly higher than the internal peer review agreement (0.40), but the correlation with the VQR journal percentile is higher (0.55). Overall, the correlation between journal metrics and peer review seems slightly lower than (or on par with) correlations among two peer reviewers at the institutional level. The citation metrics clearly show lower correlations with peer review than the journal metrics.

Table 2: Spearman correlations with reviewer 1's score at the institutional level.

GEV		Number of universities	NCS	P(top 10%)	VQR Cit. Percentile	NJS	JPP(top 10%)	VQR Journal Percentile	Reviewer 2 score
1	Mathematics and Computer Sciences	53	0.27	0.27	0.25	0.34	0.37	0.43	0.49
2	Physics	53	0.29	0.39	0.48	0.40	0.38	0.33	0.47
3	Chemistry	51	0.21	0.18	0.39	0.47	0.46	0.56	0.22
4	Earth Sciences	35	0.19	0.22	0.18	0.49	0.52	0.37	0.52
5	Biology	58	0.35	0.34	0.39	0.45	0.50	0.55	0.40
6	Medicine	47	0.36	0.44	0.57	0.48	0.46	0.66	0.59
7	Agricultural and veterinary sciences	34	0.55	0.39	0.35	0.45	0.44	0.47	0.67
8b	Civil Engineering	42	-0.10	-0.20	-0.05	0.22	0.24	0.31	0.03
9	Industrial and Information Engineering	49	0.11	0.13	0.22	0.12	0.10	0.34	0.24
11b	Psychology	30	0.18	0.10	-0.04	0.26	0.24	0.07	0.34
13	Economics and Statistics	56	0.16	0.14		0.33	0.33		0.35

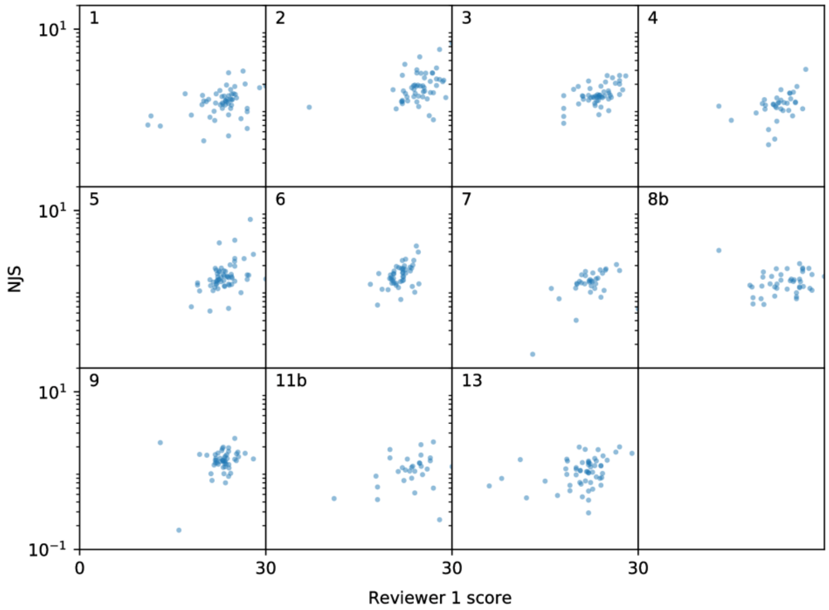

Figure 4 Scatter plot between reviewer 1's score and normalised journal score at the institutional level.

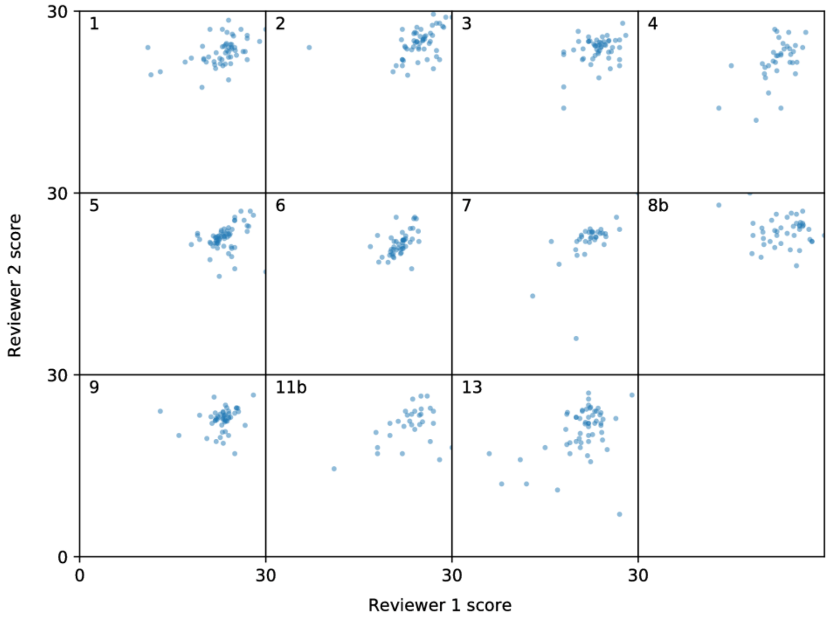

Figure 5 Scatter plot between the scores of reviewer 1 and reviewer 2 at the institutional level.

Discussion

Based on a large-scale analysis of data from the Italian VQR, we analysed the relationship between peer review and metrics at both the publication level and the institutional level. Results vary among scientific fields, but overall, we found that metrics, especially journal metrics, correlate about equally well with peer review as two peer reviewers with each other. At the publication level journal metrics correlate even a little better than two peer reviewers among each other, while at the institutional level correlations among metrics and reviews are slightly lower. In this sense, the results support, or at least do not discourage, the possibility of using metrics in combination with peer review for evaluation purposes.

We emphasise the importance of the institutional level, because this is the level at which the outcomes of research assessment exercises such as the VQR and the REF are published. The publication level is of interest because this is the most fine-grained level at which peer review and metrics can be compared. However, when outcomes are published and have consequences at the institutional level, this is the level that is most relevant from a policy point of view.

Overall, journal metrics show higher correlations with peer review than citation metrics. This may be an argument to justify the use of journal metrics to evaluate research (in line with the ideas of Waltman and Traag (2017), and in contrast to critical initiatives such as the San Francisco Declaration on Research Assessment). On the other hand, it could also be a reflection of how peer review is carried out in practice. After all, if peer reviewers largely judge a paper by the venue of publication, this would lead to a relatively strong correlation between journal metrics and peer review. Of course, if peer reviewers do not do much more than considering the venue of publication, we may also wonder what the added value of peer review is.

In order to compare metrics to peer review in an ideal setting, we would want to have peer reviews of publications independent of any characteristics such as publication venue, authors and their affiliations. Unfortunately, this is difficult to realise, since in evaluation exercises such as the VQR peer review takes place after the publication of a paper, and hence especially high impact papers are likely to be already familiar to peer reviewers.

Qualitative studies of evaluation exercises such as the VQR and the REF may provide more insight into the extent to which peer reviewers base their judgements on metrics, and in particular on journal metrics. Basically, the question is whether peer reviewers truly provide an expert assessment of the publications they review, or whether they tend to provide assessments in a semi-mechanistic way either based on easily available journal metrics or based on the reputation of journals (which is probably determined to a significant extent by journal metrics).

References

Alfò, M., Benedetto, S., Malgarini, M., & Sarlo, S. (2017). On the use of Bibliometric information for assessing articles quality : an analysis based on the third Italian research evaluation exercise. In *STI 2017. Open indicators: innovation, participation and actor-based STI Indicators*.

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. *Research Evaluation*, *24*(3), 242–255. https://doi.org/10.1093/reseval/rvv008

Anfossi, A., Ciolfi, A., Costa, F., Parisi, G., & Benedetto, S. (2016). Large-scale assessment of research outputs through a weighted combination of bibliometric indicators.

Scientometrics, *107*(2), 671–683. https://doi.org/10.1007/s11192-016-1882-9

Hicks, D. (2012). Performance-based university research funding systems. *Research Policy*, *41*(2), 251–261. https://doi.org/10.1016/j.respol.2011.09.007

Traag, V. A., & Waltman, L. (2017). Replacing peer review by metrics in the UK REF? *STI 2017. Open Indicators: Innovation, Participation and Actor-Based STI Indicators*.

Waltman, L., & Traag, V. A. (2017). Use of the journal impact factor for assessing individual articles need not be wrong. Retrieved from http://arxiv.org/abs/1703.02334

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, *5*(1), 37–47. https://doi.org/10.1016/j.joi.2010.08.001

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., … Johnson, B. (2015). *Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. *Higher Education Funding Council for England*. https://doi.org/10.13140/RG.2.1.4929.1363