Journal of Informetrics 11 (2017) 838–840

Contents lists available at ScienceDirect

Journal of Informetrics

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

Correspondence

Comments on the correspondence "On tit for tat: Franceschini and Maisano versus ANVUR regarding the Italian research assessment exercise VQR 2011–2014", J. Informetr., 11 (2017), 783–787

1. Introduction

In the following, we will briefly address the main criticisms expressed by Giovanni Abramo and Ciriaco Andrea D'Angelo in the correspondence "On tit for tat: Franceschini and Maisano versus ANVUR regarding the Italian research assessment exercise VQR 2011–2014" (Abramo & D'Angelo, 2017).

Before addressing the points they raise, however, a general note is in order to dispel a serious misconception. Abramo and D'Angelo seem to have in mind a simplistic model: there is a scientometric science that only requires to be "applied" to homogenous research universes, possibly by scientometricians or, if this unfortunately proves impossible, by "practitioners" willing to follow what scientometrics has to teach.

ANVUR's experience has shown that this model is far from realistic. ANVUR has to evaluate not only research, and not only just one kind of research, but also teaching, administrative performance, social impact, student competence etc. in a widely differentiated universe that includes—only to remain in Italy—curiosity driven and applied research in the hard sciences, scholarship in the humanities or law, art academies and musical conservatories. This evaluation needs to be tailored to specific academic, cultural, regional characteristics, taking into accounts the democratic government's choices as well as those of autonomous higher education institutions, and student interests. Evidently, this is a highly complex, and very *political* (in the noblest sense of the term) exercise.

The Italian legislator's choice not to entrust it to scientometricians, but rather to leading scientists and scholars, competent in different fields and highly familiar with the higher education system in a plurality of countries, has been indeed a wise one. And the fact that Abramo and D'Angelo are not even capable to imagine the scope of such an evaluation, even if one remains within the boundaries of "research" (itself a rather differentiated system, which requires the intimate knowledge of scientists and scholars), is highly indicative of the limited horizons within which they move.

That said, clearly scientometrics has much of value to say, and ANVUR is eager to learn from its advancement, as well as from its own mistakes, as proven by the differences between the two Research Evaluation Exercises (VQR 2004–10 and VQR 2011–14). Within the Agency, an intense debate is currently going on how to frame the VQR 2015–2019, which will take place in 2020–21. It is the first time that we have the possibility to plan in advance the Exercise and, in line with what we do in other fields of evaluation, we will find ways to benefit from the ideas and experience of the academic community, before taking our decisions.

2. Response to the main points

Let us now move to the main points raised by Abramo and D'Angelo (in press), which will be dealt with according to their order in the letter.

2.1. Section 3, third paragraph

Abramo and D'Angelo affirm that "BCG&M forgo responding to the demonstrated inefficiency in the selection of best publications (in keeping with VQR bibliometric criteria) to be submitted for evaluation". The reason why we did not mention that is simple: the "demonstrated" inefficiency was not cited by Franceschini and Maisano (2017).

DOI of the original article:http://dx.doi.org/10.1016/j.joi.2017.06.003.

http://dx.doi.org/10.1016/j.joi.2017.07.002

2.2. Section 3, fourth paragraph

The equal number of publications per researcher is the criterion adopted by the Research Assessment Exercise (RAE) and Research Excellence Framework (REF) assessment exercises in the United Kingdom, precisely those exercises that are considered by the critics of VQR as the brilliant example to follow. Abramo and D'Angelo mention the fact that "Research institutions with a higher share of professors in low publication-intensive fields are then disfavored". This statement is not correct, since in the final indices used to rank institutions the scores are normalized in each scientific area, thus almost sterilizing the effects of different averages among areas (and in order to strengthen this approach, ANVUR is now implementing standardization procedures, in order to make all main indicators random variables with zero mean and unit standard deviation). Then, Abramo and D'Angelo continue by asking the rhetorical question "Why did ANVUR then not enforce the 'equality' principle in this case, rather than offering bibliometrics for some and not others?". Even in this case, the answer is very simple: there are scientific areas, such as social sciences, law, humanities, where bibliometrics indicators are totally unreliable.

2.3. Section 3, fifth paragraph

This paragraph discusses the use of journal impact as a proxy of article impact. Abramo and D'Angelo say "We fail to understand what evaluating large or small communities has to do with the choice of the impact indicator of publications, and why 'to do so' one needs 'to evaluate relatively recent publications'." Again, the answers are straightforward. First, the appropriateness of bibliometric indicators depends on the size of the "object" to be evaluated, for instance an individual researcher or a large institution. In fact, while bibliometric indicators alone applied to the evaluation of a single researcher can fail to capture the complexity of an individual scientific personality and his/her achievements, they can be more safely applied to large samples, such as entire institutions, where the (sometimes) poor individual evaluations average out to yield smooth and reliable results. Second, the constraint of using recent publications has been enforced by the Ministerial Decree launching the VQR 2011–2014. We think that the statement "if using two indicators is more robust than using one, why not use twenty?" is rather a provocation than a comment. Indeed, we have been able to identify two relatively uncorrelated indicators, which provide better indications than a single one. Should one be able to find more uncorrelated indicators, one could use them all.

2.4. Section 3, last sentence

Here, Abramo and D'Angelo affirm that "the simple citation count is a better proxy of a publication impact than the C-J metric adopted by ANVUR". This sentence deserves a few comments.

In the self-cited paper (Abramo & D'Angelo, 2016), Abramo and D'Angelo affirm (passages in italic are our comments) "the hybrid indicator (*the one used in the VQR*) has only been recommended for citation windows of zero or 1 years (Levitt & Thelwall 2011)." Reading the paper by Levitt and Thelwall (2011), one discovers that their conclusion is rather different: "...the proposed method (*the one based only on citations*) is acceptable for articles published at least 2 years previously but for more recent articles would be disadvantageous..." So, first, zero or one year (according to Abramo and D'Angelo) was in fact two years in the cited paper, so that Abramo and D'Angelo misinterpret the citation. Second, and more important, Abramo and D'Angelo fail to consider another sentence of Levitt and Thelwall (2011) paper: "Of course, this is based on the assumption that long-term citation is a good indicator of the value of an article; some may argue that the quality of the journal in which the article is published is an equally valid indicator and hence that a similar weighted sum (*that is the VQR indicator*) should be used for all articles, irrespective of the date of publication".

We quoted the last sentence to emphasize what we intended in our comments to Franceschini and Maisano (2017) when we wrote that basing science only on "past truth" obstructs the advancement of knowledge.

In their paper Abramo, D'Angelo, and Di Costa (2010) reach a conclusion that is so straightforward that it hardly deserves 13 pages to be backed up. In essence, they make the following point. Consider two algorithms, A_t and B_t, whose outcomes depend on time, evaluated on a time frame t. Then consider the steady state outcome of algorithm A, i.e., A_∞. The authors show that A_t, when evaluated for a sufficiently large t, is a better predictor of A_∞ than B_t. From this rather trivial fact, they conclude that using only the citation (algorithm A) is better than using a composite algorithm based on citation and journal impact (algorithm B). They do not mention the implicit axiom on which all their conclusions are based, that is, the idea that A_∞ is the best indicator of the quality of an article.

It is precisely that assumption we do not accept as an indisputable truth. In a recent paper, Waltman and Traag (2017) carefully analyze the "non observable concept" of the "value of an article", and claim that the "problem of assessing an article *is* equivalent to the problem of determining the value of an article". Through an accurate theoretical analysis and extensive computer simulations they "demonstrate the possibility that the impact factor is a more accurate indicator of the value of an article than the number of citations the article has received".

To that conclusion, we add that a nation-wide research assessment exercise has, among its various by-products, that of influencing the publication habits of researchers, and, especially, of the younger ones. In this respect, including journal impact in the bibliometric indicators, with its corollary of lower acceptance rate, etc., encourages researchers to send their articles to the best journals, allowing them *inter alia* to benefit from better peer reviews. Finally, we notice that using only

citations may encourage bad practices, such as exceeding in unnecessary self-citations, citation stacking, intergroup citation practices; this is particularly true in the case of relatively recent articles, where a difference of two or three citations may significantly improve the article classification.

References

Abramo, G., & D'Angelo, C. A. (2016). Refrain from adopting the combination of citation and journal metrics to grade publications, as used in the Italian national research assessment exercise (VQR 2011–2014). *Scientometrics*, *109*(3), 2053–2065.
Abramo, G., & D'Angelo, C. A. (2017). On tit for tat: Franceschini and Maisano versus ANVUR regarding the Italian research assessment exercise VQR 2011–14. *Journal of Informetrics*, *11*, 838–840.
Abramo, G., D'Angelo, C. A., & Di Costa, F. (2010). Citations versus journal impact factor as proxy of quality: Could the latter ever be preferable? *Scientometrics*, *84*(3), 821–833.
Franceschini, F., & Maisano, D. (2017). Critical remarks on the Italian research assessment exercise VQR 2011–2014? *Journal of Informetrics*, *11*(2), 337–357.
Levitt, J. M., & Thelwall, M. (2011). A combined bibliometric indicator to predict article impact. *Information Processing and Management*, *47*(2), 300–308.
Waltman, L., & Traag, V. A. (2017). *Use of the journal impact factor for assessing individual articles need not be wrong.* arXiv.

Sergio Benedetto
Politecnico di Torino, Torino, Italy

Daniele Checchi
Andrea Graziosi*
Marco Malgarini
ANVUR, Rome, Italy

*Corresponding author.
E-mail address: andrea.graziosi@anvur.it (A. Graziosi)

1 July 2017
Available online 29 July 2017