A rejoinder to the comments of Benedetto et al. on the paper “Critical remarks on the Italian research assessment exercise VQR 2011–2014” (Journal of Informetrics, 11(2): 337–

A rejoinder to the comments of Benedetto et alii on the paper "Critical remarks on the Italian research assessment exercise VQR 2011–2014" (Journal of Informetrics, 11(2): 337-357)

Fiorenzo Franceschini[1] and Domenico Maisano[2]

[1] *fiorenzo.franceschini@polito.it* [2] *domenico.maisano@polito.it*
Politecnico di Torino, DIGEP (Department of Management and Production Engineering),
Corso Duca degli Abruzzi 24, 10129, Torino (Italy)

The paper "Critical remarks on the Italian research assessment exercise VQR 2011–2014" (Franceschini and Maisano, 2017) analyzed some vulnerabilities of the recently concluded Italian assessment exercise. Some apical (former and current) members of ANVUR promptly commented our criticisms through a letter to the editor (Benedetto et al., 2017). We believe that this letter, in spite of the rather standoffish and overconfident tone, is not very convincing. Comments are sometimes detailed and sometimes evasive, and they concern issues sometimes relevant and sometimes not: unfortunately, the comments concerning irrelevant issues are generally detailed, while those concerning the relevant ones are generally evasive.

In the following, we provide a rejoinder to the comments directed to our paper. We have chosen not to raise the tone or create controversy, limiting ourselves to discussing some of the issues addressed by Benedetto et al. (2017), in a brief and straight-to-the-point manner.

- **(Presumed) errors.** Benedetto et al. (2017) devote about one-third of their letter to illustrate seven presumed *trivial* errors (to use their expression) in our synthetic description of the VQR 2011-2014. We remark that none of these presumed trivial errors influences in any way the critical arguments developed in our paper. The curious and patient reader is invited to check this.
- **Small number of papers evaluated.** Benedetto et al. (2017) reject our critical arguments, arguing that some empirical data of the VQR 2011-2014 contradict them. Specifically, they claim the inconsistency of our example, which was aimed at showing the low discriminatory power when evaluating (bibliometrically or non-bibliometrically) the output of "more-than-decent" researchers (see page 342 of (Franceschini and Maisano, 2017)). The fact that 32.6% of the papers evaluated have been classified in class A and 63.4% in classes A or B (remember that there are five classes: A, B, C, D and E, in descending order) is nothing but a confirmation of this. Additionally, not extending the bibliometric evaluation from two (or three) papers to the totality of the papers published in the VQR period – as it would be against the "principle of treating all areas equally" – does not seem plausible (see also the recent contribution by Abramo and D'Angelo (2016)).
- **Combination of citation and journal metrics.** While remaining skeptical about the use of journal metrics at the level of individual articles, we are aware that some authors have recently

suggested not to systematically reject this possibility, at least in certain specific conditions (Waltman and Traag, 2017); the same authors encourage the scientific community to investigate this possibility with greater scientific rigor than hitherto. Despite this, the decision of ANVUR to combine citation and journal metrics in a rather naïve manner (i.e., not supported by any convincing scientific study) remains hasty (Abramo and D'Angelo, 2016). It is even hastier, when considering the high cost and the important practical implications of the VQR 2011-2014.

- **Combination of percentile ranks.** The comments of Benedetto et al. (2017) lead us to repeat that combining sub-indicators through a (weighted) sum of the relevant percentile ranks is conceptually wrong and misleading. The empirical analysis by Benedetto and Setti (2016) – which minimizes the real distorting effects of this operation – is of doubtful general validity and represents a tacit admission of their conceptual mistake. The appendix includes a pedagogical example for clarifying this aspect once again. Finally, we emphasize that whatever non-linear transformation – including the ones based on percentile ranks – inevitably distorts the interval property of the initial variables (e.g., citations and journal metrics), making the (weighted) sum of the transformed variables meaningless (Stockburger, 1996).
- **Lack of constructive attitude.** Benedetto et al. (2017) complain that our criticisms, as well as most of those so far addressed to the VQR 2004-2010 and VQR 2011-2014, are not constructive. Although our paper does not contain any detailed architectural redesign of the VQR (it was not our purpose!), we believe that it includes several hints to help ANVUR to improve or at least avoid the earlier mistakes. In over ten years of activity, ANVUR has unequivocally been revealing a lack of communication skills and listening to the Italian scientific community. These factors probably prevent ANVUR from seeing constructive aspects in the copious hints received.

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.

Benedetto, S., Setti, G. (2016) Un'analisi empirica dell'algoritmo di classificazione bibliometrica della VQR 2011-2014, available at: http://www.lavoce.info/wp-content/uploads/2016/06/algoritmo-analisi-empirica.pdf.

Benedetto, S., Checchi, D., Graziosi, A., Malgarini, M. (2017) Comments on the paper "Critical remarks on the Italian assessment exercise", Journal of Informetrics, 11 (2017), pp. 337-357. To appear on Journal of Informetrics.

Franceschini, F., Maisano, D. (2017). Critical remarks on the Italian research assessment exercise VQR 2011–2014. Journal of Informetrics, 11(2): 337-357.

Stockburger, D. (1996). Introduction to Statistics: Concepts, Models and Applications. Available online, http://www.psychstat.smsu.edu/sbk00.htm (accessed 15 March 2017).

Waltman, L., Traag, V.A. (2017). Use of the journal impact factor for assessing individual articles need not be wrong. arXiv preprint arXiv:1703.02334.

Appendix

To clarify the point relating to the (weighted) sum of percentile ranks, we present a pedagogical example, adapted from (Stockburger, 1996). Let us consider the comparison of scores obtained by

two high-school students (Suzy and Johnny) in two tests, the first one in English and the second one in Math. If the scores are distributed normally, then percentile ranks underestimate large differences in the tails of the distribution and overestimate small differences in the middle of the distribution. This is most easily understood in the illustration in Fig. A.1.

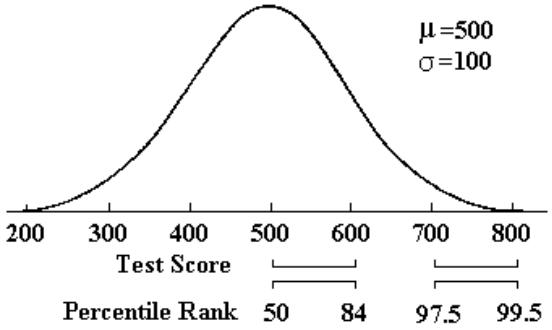

Fig. A.1. Distribution of the scores obtained in a certain test by a population of high-school students; adapted from (Stockburger, 1996).

In the above illustration, two standardized achievement tests with μ=500 and σ=100 were given. In the first one, the English test, Suzy made a score of 500 and Johnny made a score of 600, thus there was a one hundred point difference between their raw scores. In the second one, the Math test, Suzy made a score of 800 and Johnny made a score of 700, again a one hundred point difference in raw scores. It therefore can be said that the differences in the scores on the two tests were equal: one hundred points each.

When converted to percentile ranks, however, the differences are no longer equal. In the English test Suzy receives a percentile rank of 50 while Johnny gets an 84: a difference of 34 percentile rank points. On the Math test, Johnny's score is transformed to a percentile rank of 97.5 while Suzy's percentile rank is 99.5: a difference of only two percentile rank points.

It can be seen, then, that a percentile rank has a different meaning depending upon whether it occurs in the middle or the tails of the distribution; differences in the middle of the distribution are magnified, differences in the tails are minimized.

This reasoning can obviously be extended to no-matter-what other (non-uniform) distributions. The lesson learnt from this example is that not only do percentile ranks destroy the interval property, but they also destroy the information in a particular manner. Paraphrasing the concept, summing or subtracting percentile ranks is conceptually wrong and misleading.