View metadata, citation and similar papers at core.ac.uk brought to you by CORE
provided by Archivio della Ricerca - Università degli Studi di Siena

A letter on Ancaiani et. al. “Evaluating scientific research in Italy: The 2004-10 research evaluation exercise”

Published in: *Research Evaluation*, 2017

DOI: https://doi.org/10.1093/reseval/rvx013

Alberto Baccini
Department of Economics and Statistics, University of Siena, Italy
Piazza San Francesco 7, 53100 Siena, alberto.baccini@unisi.it;

Giuseppe De Nicolao,
Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy

Abstract. This letter documents some problems in Ancaiani et al. (2015). Namely the evaluation of concordance, based on Cohen's kappa, reported by Ancaiani et al. was not computed on the whole random sample of 9,199 articles, but on a subset of 7,597 articles. The kappas relative to the whole random sample were in the range 0.07–0.15, indicating an unacceptable agreement between peer review and bibliometrics. The subset was obtained by non-random exclusion of all articles for which bibliometrics produced an uncertain classification; these raw data were not disclosed, so that concordance analysis is not reproducible. The VQR-weighted kappa for Area 13 reported by Ancaiani et al. is higher than that reported by Area 13 panel and confirmed by Bertocchi et al. (2015), a difference explained by the use, under the same name, of two different set of weights. Two values of kappa reported by Ancaiani et al. differ from the corresponding ones published in the official report. Results reported by Ancaiani et al. do not support a good concordance between peer review and bibliometrics. As a consequence, the use of both techniques introduced systematic distortions in the final results of the Italian research assessment exercise. The conclusion that it is possible to use both technique as interchangeable in a research assessment exercise appears to be unsound, by being based on a misinterpretation of the statistical significance of kappa values.

KEYWORDS. informed peer review, research assessment, bibliometric evaluation, Italian VQR, peer review, Cohen’s kappa.

A letter on Ancaiani et. al. “Evaluating scientific research in Italy: The 2004-10 research evaluation exercise”.

(Ancaiani et al. 2015) described the method used in the research assessment exercise (VQR) by the Italian governmental agency for evaluation of universities and research (ANVUR). ANVUR adopted a dual system of evaluation: each piece of work was classified by informed peer review or by using bibliometric indicators. In order to validate that method, ANVUR performed a massive direct comparison between results obtained by informed peer review and by bibliometrics. Ancaiani et al. concluded that this comparison supports “the choice of using both techniques in order to assess the quality of Italian research institutions”. In this letter, some shortcomings of the comparisons are highlighted and that conclusion is challenged.

First of all, in Section 3.2, Ancaiani et al. stated that a “sample” of 9,199 articles was evaluated by both methods and results compared by using Cohen’s kappas. Truly, data of Table 2 of Ancaiani et al. do not refer to the whole random sample. As we can see in Figure 2 of Ancaiani et al., the bibliometric evaluation of papers was conducted by using a five-fold classification: four hierarchically ordered scores from A to D, plus a category IR indicating that it was impossible to use bibliometrics for deciding a score, since citations and journal impact indicators substantially diverged. If we consider instead informed peer review, each paper was evaluated by two reviewers by classifying it in one of the four categories A-D. Since it is impossible to use Cohen’s kappa for comparing results of five-category vs four-category classifications, the results reproduced in Table 2 of Ancaiani et al. cannot come from the processing of the whole 9,199-item dataset.

As a matter of fact, according to ANVUR official reports (ANVUR 2013, Appendice B), a fifth category had been defined also for peer review: “The distributions of the evaluations F and P described above are not immediately comparable, as the F distribution of the bibliometric evaluations includes a class IR that is not allowed in the peer evaluation. However, it can be hypothesized that a discordance of at least two classes between the evaluation of the first and second reviewer highlights an uncertainty of the peer evaluation analogous to that emerging from the citation number and impact factor of the publication journal in the bibliometric analysis. In analogy with the IR classification of the bibliometric evaluation, a “peer uncertain” (IP) classification was created, that allowed the comparison between the F and P2 distributions” (our translation from Italian). It is worthwhile to note that only unweighted kappa is applicable to this “5x5 protocol”: indeed for using weighted kappa it is necessary that categories are completely ordered, and in the case at hand neither peer review nor bibliometric categories can be ordered, as both include an “uncertain” category, namely IP and IR.

ANVUR published raw data for such 5x5 protocol, but, strangely enough, neither ANVUR nor Ancaiani et al. evaluated the degree of concordance on these data. Using these raw data, it is easy to verify that unweighted Cohen’s kappas for the whole sample and for areas 1-9 are in the range 0.07-0.15 (Table 1), indicating an unacceptable agreement between peer review and bibliometrics, see also Figure 1 for a graphical illustration. It is worthwhile to note that, if we consider the whole random sample, the kappas calculated for the agreement between the two reviewers (Table 2) is systematically higher that the agreement between the 5-category classifications by peer review and bibliometrics, differently from results presented by Ancaiani et. al. for a subset of the data.

Table 1. Agreement between informed peer review and bibliometrics

	Whole sample 5X5 Protocol		Reduced sample 4X4 Protocol*		
Areas	*n*	Unweighted kappa	*n*	Linear weighted kappa	VQR-weighted kappa
Area 1 Mathematics and informatics	631	0,13	438	0,32	0,32
Area 2 Physics	1412	0,12	1212	0,23	0,25
Area 3 Chemistry	927	0,14	778	0,22	0,23
Area 4 Earth Sciences	458	0,12	377	0,28	0,3
Area 5 Biology	1310	0,15	1058	0,33	0,35
Area 6 Medicine	1984	0,14	1602	0,30	0,34
Area 7 Agricoltural and veterinary sciences	532	0,12	425	0,28	0,34
Area 8a Civil Engineering	225	0,07	198	0,20	0,23
Area 9 Industrial and information engineering	1130	0,10	919	0,16	0,17
Area 13 Economics and statistics	590	0,37	590	0,54	0,61
All areas	9199	0,16	7597	0,32	0,38

Table 2. Agreement between P1 and P2

	Whole sample					Reduced sample 4X4 Protocol*		
Areas	*n*	Unweighted kappa	Linear weighted kappa	VQR-weighted kappa	VQR score	*n*	Linear weighted kappa	VQR weighted
Area 1 Mathematrics and informatics	631	0,24	0,34	0,38	0,33	438	0,36	0,35
Area 2 Physics	1412	0,15	0,24	0,28	0,23	1212	0,23	0,23
Area 3 Chemistry	927	0,15	0,22	0,26	0,21	778	0,25	0,24
Area 4 Earth Sciences	458	0,16	0,23	0,26	0,23	377	0,25	0,25
Area 5 Biology	1310	0,17	0,26	0,3	0,25	1058	0,28	0,27
Area 6 Medicine	1984	0,14	0,23	0,27	0,22	1602	0,25	0,24
Area 7 Agricoltural and veterinary sciences	532	0,07	0,17	0,2	0,17	425	0,16	0,27
Area 8a Civil Engineering	225	0,16	0,2	0,23	0,2	198	0,2	0,19
Area 9 Industrial and information engineering	1130	0,15	0,21	0,24	0,19	919	0,19	0,18
Area 13 Economics and statistics	590	0,24	0,4	0,46	0,39	590	0,4	0,46
All areas	9199	0,17	0,27	0,32	0,27	7597	0,29	0,33

Sources. *Data drawn from ANVUR report. Appendix B. Not reproducible.
All other data, our elaboration from ANVUR publicly available raw data. Appendix B of ANVUR report.
R, psyc package ver. 1.6.6 https://cran.r-project.org/web/packages/psych/psych.pdf.

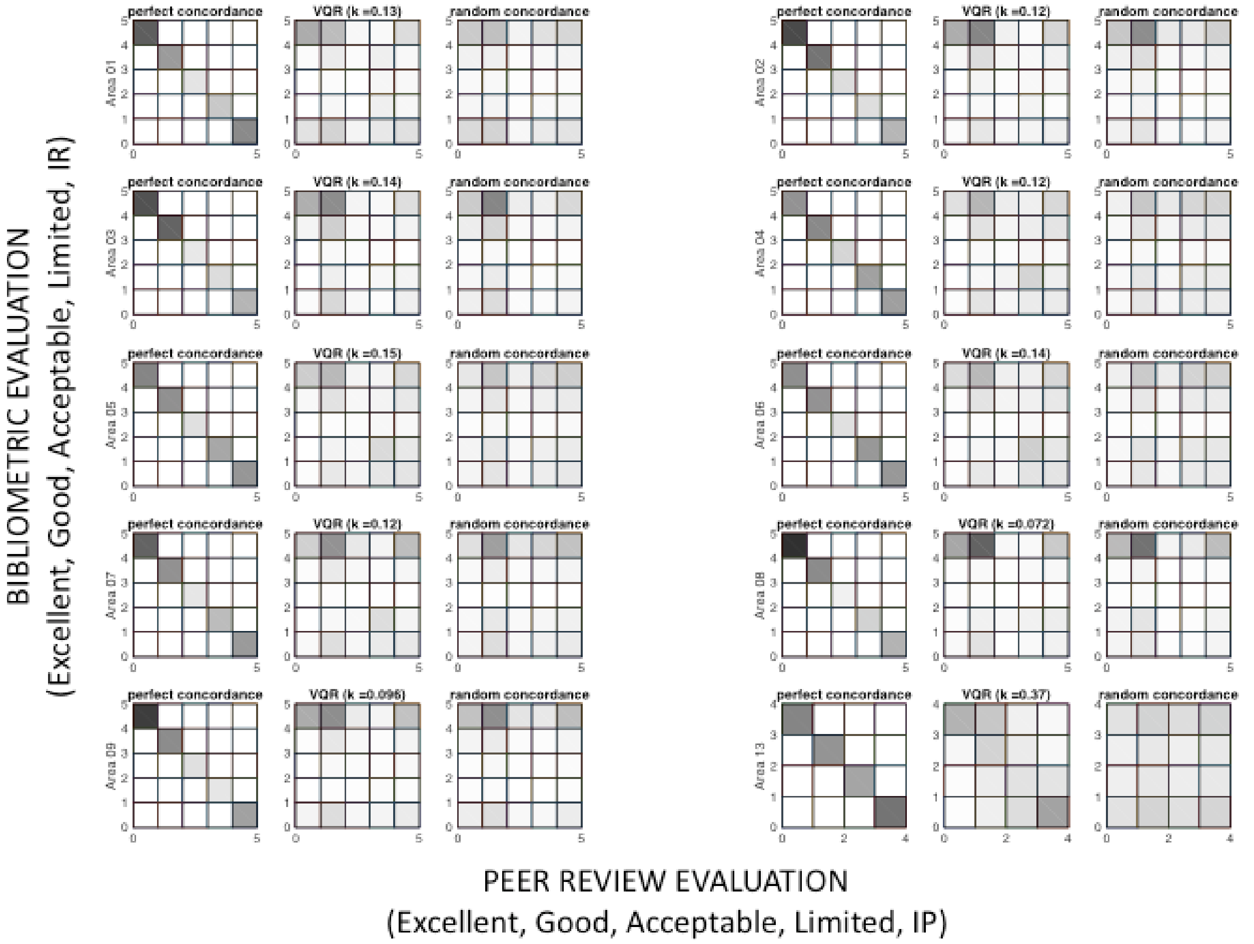


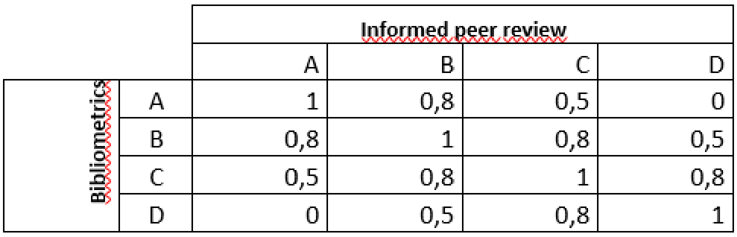
Figure 1. Ancaiani et al. state that: (i) [Cohen's] "K is always statistically different from zero, showing that there is a fundamental agreement among the two distributions"; (ii) "a significant degree of concordance among peer review and bibliometric evaluations" supports "the choice of using both techniques in order to assess the quality of Italian research institutions". However, statistical significance does not mean practical significance: according to (Cohen 1990), "The primary product of a research inquiry is one or more measures of effect size, not P values". In order to assess the size of the concordance observed in the Italian VQR, the following three joint distributions are displayed for each of the 10 areas (tones of grey are proportional to probability values). From left to right: (i) perfectly concordant distribution (kappa =1) with both marginal distributions set equal to the average of the observed VQR distributions of peer review and bibliometric evaluations; (ii) the observed VQR distribution (the unweighted kappa is reported on top); (iii) randomly concordant distribution (kappa = 0) under independent marginal distributions equal to the VQR ones. Apart from area 13, where the protocol underwent substantial changes (Baccini and De Nicolao 2016a), the observed VQR agreement (central), once compared to the extremes of perfect agreement (left) and no agreement (right), appears neither "fundamental" nor able to support the use of both techniques in order to assess research quality.

In fact, rather than being computed on the whole random sample, the values of kappa contained in Table 2 of Ancaiani et al. refer instead to a "4x4 protocol" obtained by discarding from the random sample all papers classified as IR by bibliometrics. By cross-checking with ANVUR reports, it can be seen that results presented by Ancaiani et al. actually refer to a subset made of 7,597 articles. This non-random exclusion of items from the initial random sample represents a major shortcoming of the procedure that was not disclosed by Ancaiani et al..

The adoption of the 4x4 protocol, whose categories can now be ordered, allowed the use of further indicators such as linear-weighted and VQR-weighted kappas yielding kappa values systematically higher that the unweighted ones calculated under the 5X5 protocol. Moreover, the use of a subset of the sample also for studying the agreement between the two reviewers resulted in systematic lower values of kappa with respect to those observed in the whole random sample (Table 2)

Raw data for the 4x4 protocol are not publicly available[1] and it is therefore impossible to replicate the values of kappa reported in Table 2 of Ancaiani et. al. The only exception is Area 13, where the category IR was not used at all, so that raw data for the 4X4 protocol coincide with those published in the ANVUR report. For this area we were able to replicate the estimate of k=0.6104 contained in Table 2, penultimate row, of Ancaiani et. al. (and also in ANVUR 2013: Appendix B, p. 22). Surprisingly enough, this value, representing the greatest kappa value among all scientific areas, differs from the value of k=0.54 reported in the ANVUR Area report for Area 13 and subsequently reproduced in (Bertocchi et al. 2015) and recently confirmed in (Bertocchi et al. 2016). By suitably modifying the so-called VQR-weights declared by Ancaiani et al., we were able to replicate also that second value (Table 2). We can conclude that two different sets of weights, both labelled as "VQR-weights", were actually used by ANVUR (Table 3) and the area 13 panel (Table 4). Between the two versions of VQR-weights, the results in Table 2 of Ancaiani et al. are computed using the weights that yield higher values of kappa.

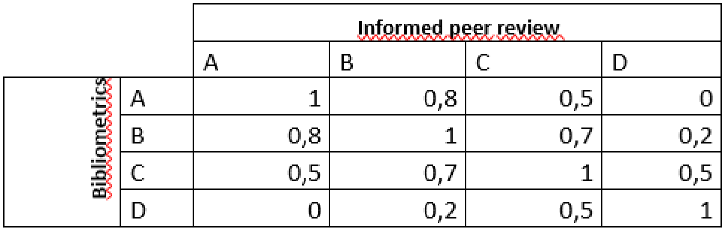
Table 3. VQR-Weights. Matrix used by ANVUR and Ancaiani et al.



		Informed peer review			
		A	B	C	D
Bibliometrics	A	1	0,8	0,5	0
	B	0,8	1	0,8	0,5
	C	0,5	0,8	1	0,8
	D	0	0,5	0,8	1

This matrix attributed to agreement, one-class, two-classes and three-classes disagreement weights modelled on the basis of the score (1, 0.8, 0.5, 0) associated to the four categories in which paper are classified (A, B, C, D). For example, consider two papers: a paper classified as A by bibliometrics and classified as B by peer-review; and a second paper classified B by bibliometrics and C by peer review. Both have a one class disagreement and a weight of 0.8, which appears arbitrary. In fact, in the former case the score error is 1.0-0.8=0.2, while in the latter one it is 0.8-0.5=0.3.

Table 4. VQR-Weights. Matrix used by Area 13 panel



		Informed peer review			
		A	B	C	D
Bibliometrics	A	1	0,8	0,5	0
	B	0,8	1	0,7	0,2
	C	0,5	0,7	1	0,5
	D	0	0,2	0,5	1

This matrix attributed to agreement, one-class, two-classes and three-classes disagreement weights modelled on the basis of the difference between the scores associated to the four categories in which papers are classified (A, B, C, D). For example, consider two papers: a paper classified as A (score 1) by bibliometrics and classified as B (score 0.8) by peer-review; and a second paper classified B (score 0.8) by bibliometrics and C (score 0.5) by peer review. Both have a one class disagreement; the difference between the two scores for the first paper is 0.2, and the weight is 1-0.2=0.8; for the second paper the difference between the two scores is 0.3, and the weight is 1-0.3=0.7.

Anyhow, the agreement between peer review and bibliometrics, as proxied by linear-weighted and VQR-weighted kappas, is still poor to fair in Areas 1-9, while the result in Area 13 is questionable due to changes introduced in the protocol with respect to the one adopted in all the other areas (Baccini and De Nicolao 2016a, 2016b). In fact, Ancaiani et al. write: "The value of K ranges from 0.16 to 0.61 depending on the area and weights, being on average equal to 0.32, a value that is usually considered as 'poor to fair'". Nevertheless, they state that "kappa is always statistically different from zero, showing that there is a *fundamental*

[1] Data had been requested to the President of the ANVUR with a mail sent the 10th February 2014. We have not received yet a reply.

agreement" among peer review and bibliometrics, claiming "evidence of a significant degree of concordance among peer review and bibliometric evaluations, supporting the choice of using both techniques in order to assess the quality of Italian research institutions". Here, Ancaiani et al. seem to mistake the notion of a parameter statistically different from zero for the notion of practical significance of its value, a well known misinterpretation, warned against in most statistical textbooks (Cohen 1990). Indeed, just as for the correlation coefficient, statistical significance "is generally of little practical value, since a relatively low value of kappa can yield a significant result. In other words, a value such as $k = 0.41$ (in spite of the fact that is statistically significant) may be deemed by a researcher to be too low a level of reliability (i.e. degree of agreement) to be utilized within a practical context" (Sheskin 2003).

In summary:

- The evaluation of concordance reported by Ancaiani et al. was not computed on the whole random sample made of 9,199 articles organized in a 5x5 protocol, but on a subset of 7,597 articles re-organized in a 4x4 protocol;
- Neither ANVUR nor Ancaiani et al. reported that the kappas relative to the whole random sample were in the range 0.07-0.15, indicating an unacceptable agreement between peer review and bibliometrics;
- The 7,597-article subset on which ANVUR and Ancaiani et al. evaluated concordance was obtained by nonrandom exclusion of all articles classified as IR by bibliometrics and, differently from the 5x5 protocol whose raw data are in the VQR report, ANVUR did not disclose raw data, so that concordance analysis is not reproducible with the exception of Area 13;
- The VQR weighted kappa for Area 13 reported by ANVUR and by Ancaiani et al. is higher than that reported by Area 13 panel and confirmed by Bertocchi et al., a difference explained by the use of two different set of weights, both labelled as "VQR weights";
- Two values of kappa reported by Ancaiani et al. in Table 2, differ from the corresponding ones published in the ANVUR report;
- When claiming that their results support the interchangeable use of bibliometrics and peer review in a research assessment, Ancaiani et al. seem to confound the notion of a kappa statistically different from zero with that of practical significance of its value, which, according to literature guidelines, is "poor to fair".

In short, despite all efforts, the results reported by Ancaiani et al. do not support a good concordance between peer review and bibliometrics. As a consequence, the use of both techniques introduced systematic distortions in the final results of the Italian research assessment exercise. On the basis of these data, the conclusion that it is possible to use both technique as interchangeable in a research assessment exercise appears to be unsound.

REFERENCES

Ancaiani, Alessio, et al. (2015), 'Evaluating scientific research in Italy: The 2004–10 research evaluation exercise', *Research Evaluation,* 24 (3), 242-55.

ANVUR (2013), 'Rapporto finale. Valutazione della qualità della ricerca 2004-2010 (VQR 2004-2010)', (Roma: http://www.anvur.org/rapporto/).

Baccini, Alberto and De Nicolao, Giuseppe (2016a), 'Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise', *Scientometrics,* 108 (3), 1651-71.

--- (2016b), 'Reply to the comment of Bertocchi et al', *Scientometrics,* 108 (3), 1675-84.

Bertocchi, Graziella, et al. (2015), 'Bibliometric evaluation vs. informed peer review: Evidence from Italy', *Research Policy,* 44 (2), 451-66.

Bertocchi, Graziella, et al. (2016), 'Comment to: Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise', *Scientometrics*, 1-5.

Cohen, Jacob (1990), 'Things I have learned (so far)', *American Psychologist,* 45 (12), 1304-12.

Sheskin, David J. (2003), *Handbook of Parametric and Nonparametric Statistical Procedures* (London: Chapman & Hall).