Peer review and bibliometric indicators just don't match up according to re-analysis of Italian research evaluation.

blogs.lse.ac.uk/impactofsocialsciences/2016/06/16/peer-review-and-bibliometric-indicators-just-dont-agree/

The Italian research evaluation agency undertook an extensive analysis to compare the results of peer review and bibliometric indicators for research evaluation. Their findings suggested both indicators produced similar results. Researchers **Alberto Baccini** *and* **Giuseppe De Nicolao** *re-examine these results and find notable disagreements between the two techniques of evaluation in the sample and outline below the major shortcoming in the Italian Agency's interpretation. Results from one technique will differ from those reached using the other.*

> *"Individual metrics give significantly different outcomes from the REF [Research Excellence Framework] peer review process, showing that metrics cannot provide a like-for-like replacement for REF peer review". – HEFCE (2015). The Metric Tide: Correlation analysis of REF2014 scores and metrics (Supplementary Report II)*

This sentence summarizes results of the comparison between REF 2014 peer review and many metric indicators performed by the "Independent review of the role of metrics in research assessment and management". This overwhelming evidence is apparently challenged by results obtained in the Italian research assessment exercise (VQR 2004-2010), managed by the Italian Agency for the evaluation of the university and research (ANVUR) and concluded in 2013. During this assessment a large comparison between bibliometric evaluation and peer review was performed and its results were summarized as showing that bibliometric analysis and peer review produced "very similar evaluation". The policy conclusion is therefore drawn that the two techniques of evaluation can be used contemporaneously in research assessment since they have to be considered as substitute. In this post, based on an article recently published in *Scientometrics*, we argue that if appropriate statistical guidelines are adopted for analyzing data, also in the Italian experiment, peer review and bibliometrics don't agree.

Peer review and bibliometric in the Italian research assessment exercise

The Italian research assessment exercise adopted a "dual system of evaluation": each research output submitted was classified in one of four merit classes (Excellent, Good, Acceptable, Limited) by informed peer review or by using bibliometric indicators. For informed peer review, members of research field panels asked two anonymous referees to evaluate articles by providing them with metadata of the articles and bibliometric indcators. Each referee produced a score, and the two scores were finally summarized in a final evaluation. The bibliometric evaluation was assigned to each article according to involved field specific criteria, based on the total number of citation received by an article and the impact factor of the journal in which the article had been published.

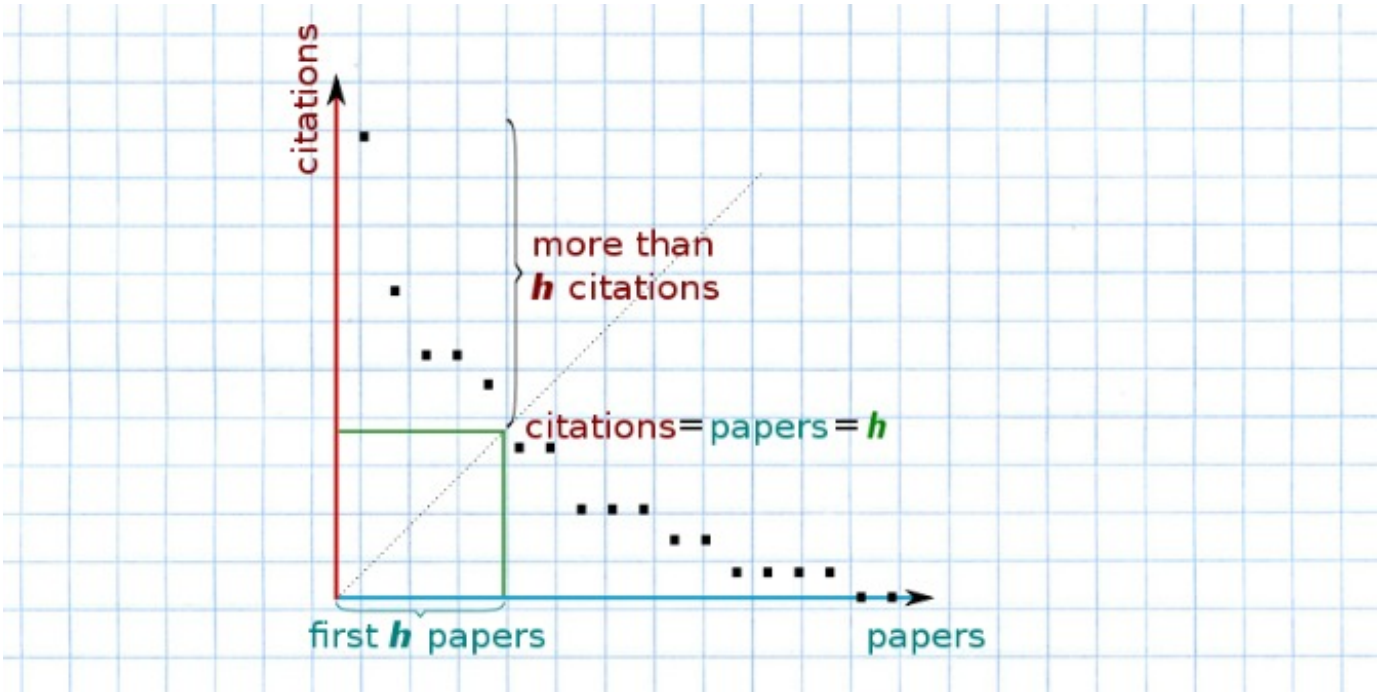

Image credit: *h*-index from a plot of decreasing citations for numbered papers (public domain, wikipedia)

In order to validate the use of this dual system, ANVUR performed an experiment of massive comparison among peer review and bibliometric indicators. Indeed, 9,199 journal articles submitted for the assessment were evaluated by ANVUR by using both bibliometrics and peer review. These articles represented the 9.3% of the total number of the journal articles submitted to VQR. All fields of hard sciences, biological sciences, medicine, engineering and the so called Area 13 (economics, management and statistics) were included in the experiment.

The bulk of the experiment consisted in the analysis of the agreement between the evaluations obtained by using the two methods of evaluation. For this comparison, ANVUR adopted Cohen's kappa, the most popular index of interrater agreement for nominal categories. The upper limit for kappa value is 1 occurring when bibliometrics and peer review perfectly agree; values of kappa near to zero indicates that observed agreement is less than or equal to the agreement expected by chance. Cohen's kappa were computed in two different ways by using linear weights and the so-called "VQR-weights". For the whole experiment, the first technique gave a value of 0.32 and the second of 0.38: in the whole sample of more than nine thousands articles, just a little less or a little more than one third of evaluations reached through bibliometrics and peer review are in agreement.

Kappas were also computed separately for 10 research fields; they are all in the range 0.1-0.35, with the only exception of a value of 0.54 for Area 13. When the 43 sub-fields are considered, the minimum kappa was calculated for electronic engineering (0.09) and the maximum one for economics (0.56); and again only four sub-areas have kappa values greater than 0.40.

ANVUR 's *Final Report* described generally these data as indicating "a *good* degree of agreement for the whole sample and for each GEV" [italics added]. ANVUR summarized results of the experiment by stating that "In the complex … there is *a more than adequate concordance* between evaluation carried out through peer review and through bibliometrics. This result fully justifies the choice made at VQR […] to use both techniques of assessment". These data and their interpretation were largely disseminated by ANVUR and its collaborators not only as working papers and columns in policy blogs, but also as papers in scholarly journals.

It is therefore a bit surprisingly that a major shortcoming in ANVUR interpretation was unobserved by so many reviewers and editors. Indeed, when the existing guidelines for interpreting Cohen's kappa (available in our

Scientometrics paper) are considered, almost all the values reached in the experiment have to be considered as indicative of agreement that would be deemed as “unacceptable”, or alternatively as “poor” or “fair”. Hence, contrary to ANVUR’s claim, a correct reading of the experiment results indicates that the hypothesis of agreement between peer review and bibliometrics is not fulfilled.

Economics, statistics and management: a fatally flawed experiment.

But, there is an exception: results for Area 13. Results for economics, management and statistics showed an agreement that can be described as “acceptable” or alternatively as “fair to good” or “moderate”. A meta-analysis of the experiment confirmed that results for Area 13 (one area out of 10) and its sub-fields (three out of 43), were statistically different from the results reached for all other fields and sub-fields, as reported in Figure 1.

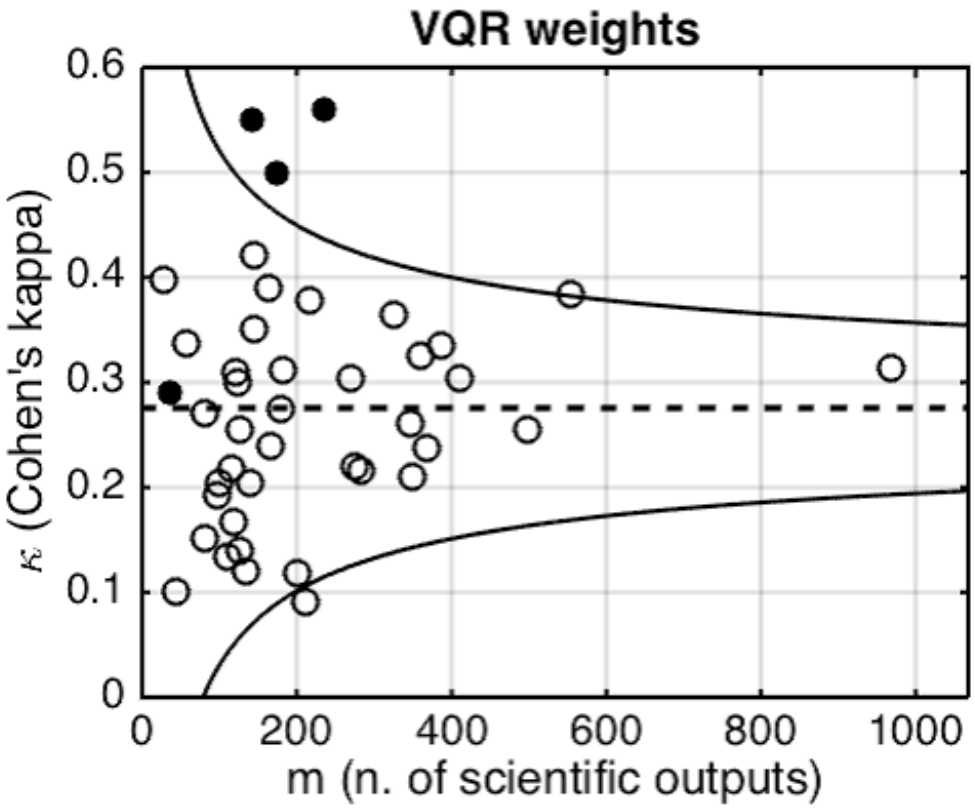

Figure 1. Funnel plot relative to the meta-analysis of the VQR experiment. A point with coordinates (m, kappa) represents a subarea having m evaluated products and whose Cohen’s agreement rate is kappa. Cohen’s kappas for Area 13-Economic and Statistical Sciences (full circles) are compared to the mean kappa (dashed) and 95 % prediction limits (continuous), based on subarea kappas collected in the other areas (open circles). In Area 13, three out of four subareas fall outside the 95% prediction limits. In the plot, VQR-weighted kappas are considered.

Why does it happen? Our response to this question, based on a careful scrutiny of the official reports, is that the higher level of agreement between bibliometrics and peer review may be the result of the modifications introduced by the Area 13 panel to the experiment protocol adopted in all the other research fields.

First and foremost, Area 13 peer reviewers, differently from the reviewers of all other research fields, knew that they were participating to the experiment. They knew also the bibliometric evaluation with which their judgement was to be compared. Indeed in Area 13 bibliometric evaluation was based on journal rankings published by the area panel. In all the other research fields, bibliometric criteria, as we have anticipated, were far more complex by involving both citation and impact factor percentiles, making the bibliometric classification much more difficult to be foreseen

by referees.

Other relevant modifications related to the way in which the synthesis of referees' report was reached. In all the other fields, final evaluation was algoritmically computed on the basis of referees' scores. In Area 13 instead, the final peer review classification of an article was decided in many cases by a "consensus group" formed by two panel members. They directly decided the final evaluation of the article, by considering the referee's reports as simple information. Under weak assumptions, it results that at least 326 articles out of 590 (55,3%) considered for the experiment were evaluated not by referees, but by the consensus groups, that is by members of the panel, by obviously knowing that they were partecipating to the experiment.

This knowledge introduced another anomaly in the experiment protocol. Since panel members of Area 13 charged to choose referees for an article knew also the merit class assigned to it by bibliometric evaluation, they might have piloted the fate of an article through a suitable choice of referees. Consider for example an article published in a journal classified as excellent, which applied standard techniques to a standard problem. If a GEV member desired to raise the probability that peer review agreed with bibliometric evaluation, she had to choose a referee who notoriously appreciated that standard technique.

We can conclude that results of the experiment for Area 13, the only research field showing a moderate agreement between peer review and bibliometrics, appears to be fatally flawed.

Conclusions

A careful analysis of the Italian experiment highlights a poor agreement between peer review and bibliometrics and supports the conclusion that they do not produce similar results. As a consequence the adoption of the dual system of evaluation in the Italian research assessment exercise possibly introduced systematic and unknown biases in its final results. Since peer review resulted in systematically lower scores, it is impossible to affirm, for example, in a comparison between two departments, if a department has a higher score because it had produced better research or because it was evaluated with a different mix of peer review and bibliometrics.

This suggest a policy conclusion opposite to the one endorsed by ANVUR: agencies that run research assessments should expect that results that will be reached by using one technique will differ from those reached using the other. A result more coherent with previous literature and with evidence provided for the REF by the correlation analysis performed in the context of the "Independent review of the role of metrics in research assessment and management".

This blogpost is based on the journal article Baccini, A. and G. De Nicolao, Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise. Scientometrics, 2016: p. 1-21.

Featured image: Anne-Lise Heinrich mismatch CC BY

Note: This article gives the views of the author, and not the position of the LSE Impact blog, nor of the London School of Economics. Please review our Comments Policy if you have any concerns on posting a comment below.

About the Authors

Alberto Baccini *is Professor of Economics in the Department of Economics and Statistics, University of Siena. Italy*

Giuseppe De Nicolao *is Professor in the Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy*