# Reply to the comment of Bertocchi et al.

# Reply to the comment of Bertocchi et al.

Alberto Baccini
(Dept. of Economics and Statistics, University of Siena. Italy
Piazza San Francesco 7, 53100 Siena, alberto.baccini@unisi.it; tel. +390577233076)

Giuseppe De Nicolao
(Dept. of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy)

**ABSTRACT.** The aim of this note is to reply to Bertocchi et al.'s comment to our paper "Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise". Our paper analyzed results of the experiment conducted by the Italian governmental agency ANVUR during the research assessment exercise about the agreement between informed peer review (IR) and bibliometrics. We argued that according to available statistical guidelines, results of the experiment are indicative of a poor agreement in all research fields with only one exception, results reached in the so called Area 13 (economics and statistics). We argued that this difference was due to the changes introduced in Area 13 with respect to the protocol adopted in all the other areas. Bertocchi et al.'s comment dismiss our explanation and suggest that the difference was due to "differences in the evaluation processes between Area 13 and other areas". In addition, they state that all our five claims about Area 13 experiment protocol "are either incorrect or not based on any evidence". Based on textual evidence drawn from ANVUR official reports, we show that: (i) none of the four differences listed by Bertocchi et al. is peculiar of Area 13; (ii) their five arguments contesting our claims about the experiment protocol are all contradicted by official records of the experiment itself.

*Scientometrics*

published on line 09 July 2016

DOI 10.1007/s11192-016-2055-6

**Keywords:** Bibliometrics; Informed peer review; research assessment; bibliometric evaluation; meta-analysis; peer review; Italian VQR; open science

The aim of this note is to reply to the comment of Bertocchi et al. (2016) comment to our paper (Alberto Baccini and De Nicolao 2016): "Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise". Our paper analyzed results of the experiment conducted by the Italian governmental agency ANVUR during the research assessment exercise (VQR) about the agreement between informed peer review (IR) and bibliometrics. We argued that according to available statistical guidelines, results of the experiment have to be considered as indicative of a poor agreement in all research fields with only one exception, results reached in the so called Area 13 (economics and statistics). Results of Area 13 were also publicized as an article published by *Research Policy* (Bertocchi et al. 2015) that we used as a reference for a part of our paper.

Bertocchi et al.'s comment agreed with us on one of the central points of our paper: results of the experiment for area 13 have to be considered as very different from the ones reached for all the other areas. They don't agree with us about the probable causes of this difference. We argued that this difference was due to the changes introduced in Area 13 with respect to the protocol adopted in all the other areas. They dismiss our explanation and mention "differences in the evaluation processes between Area 13 and other areas" (p. 2), listing four differences (p. 2) in the evaluation processes. In addition, they state that our all five claims about Area 13 experiment protocol "are either incorrect or not based on any evidence" (p. 3).

Based on textual evidence drawn from ANVUR official reports, we show that: (i) none of the four differences listed by Bertocchi et al. is peculiar of Area 13; (ii) their five arguments contesting our claims about the experiment protocol are all contradicted by official records of the experiment itself.

In the following, a point by point rebuttal of Bertocchi et al.'s is given. Our statements are supported by detailed textual analyses of the sources.

**Was Area 13 the white raven of Italian research assessment exercise?**

In the section entitled *The Italian research assessment exercise* of our paper we summarily documented the differences among assessment techniques adopted in all research areas in the VQR: Area 13 is but one of 10 different research assessments conducted by using different techniques.[1] As a consequence, also the experiment about agreement between IR and bibliometrics can be considered as composed by ten different experiments. By starting from this rationale, we developed a meta-analysis for searching for a common truth behind these ten different experiments.

Therefore, the point raised by Bertocchi et al. that Area 13 adopted different procedures for the assessment is not relevant to our analysis, because our starting point was exactly that all the 10 areas had used different assessment techniques. Bertocchi et al.'s comments seems to suggest that Area 13 was a sort of white raven of the Italian research assessment, adopting a peculiar technique of assessment, whereas all the other areas adopted a common assessment technique. But this was not the case.

In particular none of the four differences listed at p. 2 of Bertocchi et al.'s comment are peculiar for Area 13:

1. "In all other areas, researchers knew in advance the journal and the citation classification system, while in Area 13 the ranking of journals was published only after authors submitted their papers."

[1] We agree with Bertocchi et al. that final results of the assessment cannot be compared, as we clearly stated on p. 6 of our paper: "there was lack of comparability not only between Areas but also between research fields inside the same research Area". It is therefore a bit disappointing to read on p. 2 of Bertocchi et al.'s comment that "A crucial point of the VQR is that evaluations cannot be compared directly across research areas (which differ in terms of publication standards, publication types, refereeing style, citations, etc.). The entire BD paper is instead based on such comparison." The absence in our paper of any comparison of results of the research assessment is something that can be verified very easily, given that we did not present any result of the research assessment at all.

This claim is contradicted by ANVUR official records. As detailed in the Appendix A, information available about journal rankings and citation thresholds for researchers before submission was largely incomplete for all the ten Areas considered and Area 13 cannot be considered as an exception.

2. "The panel evaluating Area 13 based the classification of journals on a combination of 5-year Impact Factor (5IF), 5-year Article Influence Score (AIS), and citation analysis. 5IF and AIS are arguably more robust measures than the simple Impact Factor, which was the bibliometric indicator used in all other areas."
   Also this claim does not correspond to ANVUR official records. Indeed, as we wrote in our paper, many journal bibliometric indicators were used in the VQR. For example, in Area 9 journals were classified by using a principal component analysis applied to impact factor, 5IF, AIS and eigenfactor score; Area 1 used data from three different sources (WoS, Scopus and Mathscinet), by using also a specific algorithm for ranking journals. A possibly complete description is available in (A. Baccini 2016). In short, there were many differences in bibliometric assessment tools used by all the different research areas, and Area 13 cannot be considered as a specific case.

3. "In Area 13 the weight of citations in the bibliometric classification was different than in other areas (for instance, there were no ''downgrades'' for journal articles with few citations)."
   As summarily explained in our paper, all ten areas used different weights for citations: citations and their threshold varied among areas. Also in this case, the "no downgrading" rule was adopted not only by the area 13 panel, but at least also by the Area 1 panel.[2]

4. "Area 13 included in the journal list also journals not included in WoS, by an imputation method described in the RP paper."
   Also in this case, the use of an imputation method for ranking journals not included in WoS was not peculiar to Area 13 alone. Indeed, Area 08 used a ranking of journals containing both WoS indexed journals and journals classified according to criteria defined by the panel; and in Area 1, for some subjects, the list of WoS journals was integrated by journals included in MathSciNet.[3]

In short, in the Italian research assessment exercise each Area adopted specific rules for assessment. Area 13 was only one of ten different cases, not a particular case against other nine areas adopting a common assessment method. Moreover, neither in (Bertocchi et al. 2015) nor in Bertocchi et al's. comment, it is explained how the alleged assessment peculiarities of Area 13 might have induced a better agreement, thus affecting the Cohen's kappas of the experiment.

**Again on the changes of the experiment protocol for Area 13**

Our paper argued that modifications to the experiment protocol introduced by Area 13 panel were responsible for the higher degree of agreement registered in Area 13. We listed five modifications that Bertocchi et al. considered as "either incorrect or not based on any evidence". In what follow we will show that none of the five objections made by Bertocchi et al. in their comment is supported by ANVUR official reports; we will show also that all our five claims are clearly and soundly documented in the ANVUR official reports.

---

[2] "Citation may improve the merit class than that of the journal in which the article is published, or at least lead to a more detailed examination of the article in case of significant difference between the merti class of the journal and the indicator given by the number of citations". Faq n. 13 in http://www.anvur.org/attachments/article/77/gev01_faq.pdf.

[3] http://www.anvur.org/rapporto/files/Area01/VQR2004-2010_Area01_RapportoFinale.pdf; http://www.anvur.org/rapporto/files/Area08/VQR2004-2010_Area08_RapportoFinale.pdf.

1. Our first claim regarding Area 13 experiment was that: "the random sampling of articles in Area 13 was possibly affected by authors's requests of being evaluated through peer review. No information about the extent of these insertions is currently available" (p. 16Alberto Baccini and De Nicolao 2016) This claim originated, as we wrote in note 22, from a statement on p. 64 of the Appendix B of the Area 13 Report, that was dropped in (Bertocchi et al. 2015):

   > "The sample selection shall take account of any specific request for peer review reported via the CINECA electronic form for highly specialized and multidisciplinary products".[4]

   Area 13 panel is the only panel that introduced this description in the definition of experiment sample. Now, Bertocchi et al. 's comment (p. 3) confirms our claim about departure from random sampling;

   > "The panel received some requests for peer review, but they referred *mostly* to ''multidisciplinary'' papers, which—by the rules of the VQR—were evaluated jointly with other panels by peer review. So requests for peer review did not affect the sampling of journal articles. [italic added]".

   Just as before, data about the process are not disclosed, nor is motivated this departure from the protocol of all the other areas. We are left to wonder how many papers were in the "mostly" group and how many were in the "non-mostly" one. Finally, we are left to wonder why the panel introduced this modication to the simple random sampling adopted by all the other panels.

2. Bertocchi et . al. wrote: "A second claim [of BD] is that panel members and reviewers knew that a journal article sent in peer review was part of the experiment and the bibliometric evaluation of the article." In particular, they object that the list of journal was incomplete and that 1,135 journal articles were evaluated by peer review.

   We confirm our claim. Panel members, reviewers and researchers in Area 13 knew that journal articles published in the journals ranked by the panel, *unless included in the experiment,* would be evaluated by bibliometrics alone, that is by considering the rank of the journal. Indeed, in the evaluation criteria we can read: "The GEV will evaluate all journal articles by bibliometric analysis, and at least the 10% of these articles also by peer review".[5] The definitive ranking of journals was published before the beginning of peer evaluation process.[6] The score deriving from journal ranking would be upgraded of one class if the articles had received more at least five citations per year.[7] Moreover, as finally disclosed in (Bertocchi et al. 2015): "The referees were provided with the panel journal classification list".

   It was straightforward for panel members and reviewers to infer that all articles they received and for which the journal classification was provided, were included in the sample of the experiment. Conversely, it was straightforward for panel members and reviewers that articles for which no journal classification was provided, were not included in the sample of the experiment.

---

[4] http://www.anvur.org/rapporto/files/Area13/VQR2004-2010_Area13_Appendici.pdf

[5] "The [panel] will evaluate all the journal articles by bibliometrics, and at least the 10% of the same articles also by peer review" [translation by the authors];
http://www.anvur.org/attachments/article/92/gev13_criteri.pdf, p. 3.

[6] "Starting from that date [4th September 2012, the date of publication of the final journal ranking], preceding the beginning of the peer review evaluation, the list was not integrated or corrected" (p. 5 of http://www.anvur.org/rapporto/files/Area13/VQR2004-2010_Area13_RapportoFinale.pdf) and again: "Peer review evaluation took place from the end of September 2012 to Febraury 2013" (p. 116 of http://www.anvur.org/rapporto/files/Area13/VQR2004-2010_Area13_Appendici.pdf). [Translation by the authors].

[7] This is stated on the panel official document dated of 2nd april 2012 available here: http://www.anvur.org/attachments/article/92/gev13_allegati.zip

It is therefore absolutely irrelevant that journal ranking was incomplete. And it is also irrelevant to know the, likely wrong,[8] total number of journal articles sent to peer reviewers that Bertocchi et. al. provided in their comment.

It is instead worthwhile to stress again that the state of information of peer reviewers in Area 13 was unique in the experiment. In the other areas, peer reviewers did not know if they were participating to the experiment because it was impossible to for them to distinguish between an article sent for the experiment and the other ones. To be definitively clear: consider a reviewer in Area 13 receiving an article published in a journal listed in the journal ranking; she immediately knew that she was participating to the experiment because the article according to the panel criteria had to be evaluated by bibliometrics. In all the other areas, when a reviewer received an article published in a journal listed in the Web of Science (or in one of the other lists available), she did not know if she was participating to the experiment because also journal papers for which the bibliometric algorithm did not provide a definitive score were sent in peer review.

3. Bertocchi et al.'s comment attributes to us a third claim that we never made: "the fact that the experiment did not compare anonymous manuscripts in peer review with bibliometric indicators, and that referees knew the ranking of journals."

   In our paper it is stated clearly that discussion is about informed peer review, that is, peer review in which reviewers are provided with complete metadata of the article and bibliometric indicators. This was not peculiar to Area 13. What was peculiar to Area 13 was that, as seen in point 2, the reviewers knew also the bibliometric evaluation with which their judgement would eventually be compared. This information was not available to reviewers of all the other areas for the reasons already discussed in our paper, and not disputed in Bertocchi et al.'s comment.

4. Bertocchi et al. contest our fourth claim, by presenting a description of the protocols used for synthesizing reviewers reports that is at odds with ANVUR official reports. The issue at hand is how final peer review scores for each article were synthesized from two referees reports. In our report we claim that at least 326 articles out of 590 were evaluated by Consensus Groups. Bertocchi et al. comment that there were at most 15 such cases.

   First of all, we confirm our claim that, according to ANVUR official area reports, the protocol used for synthesizing referees' scores in the Area 13, was different from the common protocol adopted by all the other areas. In further support of this claim, in Appendix B we provide a description of the experiment protocol that adds some details to the one contained in our paper.

   Second: how many Consensus groups were formed? According to ANVUR official reports, Area 13 panel evaluated by peer review a total of 6.277 research outputs, by including also the 590 journal articles of the sample.[9] The panel "proceeded to the synthetic evaluation by constituting, 6.277 consensus group, one for each evaluated product".[10] According to ANVUR official reports, we can infer that for each of the 590 journal articles of the sample a consensus group was formed.[11]

---

[8] Data provided in their comments are at odds with data published in ANVUR final reports. They wrote: "The panel received 6,816 journal articles for evaluation". The final reports of Area 13 instead recorded 7,457 journal articles (Table 1.7 and 2.5) submitted and 7,442 evaluated by bibliometrics or peer review (Table 5.5) http://www.anvur.org/rapporto/files/Area13/VQR2004-2010_Area13_RapportoFinale.pdf.

[9] *ibidem*, p. 60.

[10] *Ibidem*, p. 64. Translation by the authors.

[11] For highlighting the procedural differences between Area 13 and all the other Areas, it is useful to note that a couples of area reports disclosed the total number of consensus group activated. According to the Area 7 report "938 Consensus groups were activated out of 9.878 evaluated products" http://www.anvur.org/rapporto/files/Area07/VQR2004-2010_Area07_RapportoFinale.pdf, p. 31; analogously in Area 9, 1,610 consensus groups were activated out of 7,500 research products evaluated , http://www.anvur.org/rapporto/files/Area09/VQR2004-2010_Area09_RapportoFinale.pdf , p. 20.

Third: how many articles were eventually scored by a consensus group? According to the ANVUR final report, the two referees agreed in their evaluation for 264 journal articles out of 590. In these 264 case we can suppose that Consensus groups did not modify the agreed score of the two reviewers. Reviewers disagreed for the remaining 326 journal articles, for which the final score was decided by the consensus group, according to the procedure described in Appendix B.[12]
We conclude that the statement by Bertocchi et al. that "at most 15 papers (not 326) were evaluated by the panel itself", does not correspond to information and data contained in ANVUR official reports.[13]

5. In their fifth point, Bertocchi et al.'s allegation about our claim is simply false. They write that "BD hint that panel members coordinated with referees to increase the agreement between bibliometric evaluation and peer review". Indeed, we asserted a very different thing, that we repeat here for the sake of clarity: the modifications introduced in Area 13 to the protocol of the experiment gave information to the panelists in charge of choosing reviewers that were not available to panelists of the other areas.[15] This information represented a major problem in the design of the experiment , given that this modified protocol could not rule out an opportunistic choice of the reviewers by the panelists. We did not affirm that this happened, because evidence is not available. We just stated that a social experiment whose final results rely on a hypothesis of ethical behavior of the subjects is a poorly designed experiment. It is somehow surprising that Bertocchi et al. are unaware of such a weakness, given that economists are used to models in which rational agents act opportunistically if they have the interest and the occasion to do so.

**Concluding remarks**

In sum, the modifications of the protocol introduced in Area 13 possibly introduced a bias toward agreement between bibliometrics and informed peer review. These modifications were neither explicitly highlighted in ANVUR reports, nor disclosed in (Bertocchi et al. 2015). Moreover, the introduction of these modification in Area 13 protocol with respect to the protocol adopted in all the other areas was not justified at all. We think that a generic statement about a specificity of economics and statistics with respect to the other hard sciences is not sufficient to justify these modifications.
Area 13 panelists developed the ranking of journals for bibliometric evaluation; they chose the reviewers; they formed the consensus groups that decided the final score of the journals articles when the two reviewers did not agree in their judgement. Reviewers of Area 13 knew the ranking of journals and knew that they were participating to the experiment. It is hardly surprising that in Area 13 the agreement between bibliometric and peer review evaluation reached a level not recorded in all the other areas. This kind of result cannot be considered as a sound premise for drawing policy conclusions.[16]

---

[12] It is not relevant that the choices of consensus groups were partially constrained. The existence of a constrained procedure and of a web platform, to the best of our knowledge, was disclosed only in the Bertocchi et al.'s comment.

[13] The number 15 refers to the number of articles of the experiment for which a three-class disagreement between reviewers was registered. Bertocchi et al. indicate a wrong table (Table 11) as the source of this datum. It can be found on Table 12 of (Bertocchi et al. 2015).

[15] Analogously, they wrote that we made "the allegation that [Area 13] panel has manipulated the data". Also this allegation is false. The wording "manipulated experiment" is a technical expression that indicates that the conditions of an experiment were deliberately changed by experimenters. In this case the expression simply means that the protocol of the Area 13 experiment was deliberately modified by the panel.

[16] (Bertocchi et al. 2015) were aware of at least one of the problems of the experiment: "the influence exerted on the reviewers by the information on the publication outlet implies that, in our study, assessment by bibliometric analysis and peer review are not independent". If this is true, it is really difficult to understand how they could draw the following policy conclusion: "the agencies that run these evaluations could feel confident

Our final remark is on the asymmetry of this scientific debate. A scientific debate in which data are not made available to scholars for controlling and reproducing results is a bit surreal. In fact, Anvur never replied to our request to access raw data for study purposes. As a consequence, when writing our original paper and this reply we were forced to rely on research reports and tables without having access to raw data. Consider the discussion of point 4 above. A definitive answer could be simply obtained by accessing raw data and verifying if and how consensus groups modified reviewers scores. We are afraid that we are still a long time away from the widespread adoption an open data policy by either scholars - writing papers only if data are publicly distributable - or journals - accepting articles only if raw data are publicly available.

about using bibliometric evaluations and interpret the results as highly correlated with what they would obtain if they performed informed peer review".

**Appendix A: availability of journal and the citation classification system**

The deadline for submission was 15th June 2012. For Area 13 a first provisional list of journals was published the 30th April 2012 containing data about two years impact factor, five years impact factor, and *h*-index. Updates to the list were published the 10th May and the 12th June 2012. The final ranking was published on 4th September 2012.[17] Moreover, the 2ndApril 2012 the GEV published "for better steering professors' choices" the citation classification system, in this case a simple threshold of five citations per year: if a journal article had received five of more citation per year it will be automatically inserted in the merit class immediately superior to the one of the journal in which it was published.[18]

The situation of Area 13 was not so different from the ones of all the other areas. Indeed Area 1 and Area 9 published provisional lists of journals to be integrated after the completion of the submission process.[19] Areas 2,3,4,7 did not publish list of journals by limiting to refer to the *Journal of Citation Reports* (JCR), where professors had to consider the proper distribution of journals in the relevant subject categories and years. Area 7 published also a list of journals included in the JCR for which bibliometric classification did not apply.[20] Areas 5 and 6 did not publish a list of journals, by referring to JCS again, but in these areas the subject categories were defined by joining-up groups of Web of Science subject categories. No areas published in advance citation classification systems that were going to be used for articles evaluation.[21]

**Appendix B: protocol used for synthesizing reviewers reports**

ANVUR final reports summarized this part of the procedure as follows:

> "The reviewers's evaluations were then synthesized in a final evaluation on the basis of algorithms specifically defined by each Area panel, and described in details in the Area reports" [22]

[17] These data are drawn from pp. 4-5 of http://www.anvur.org/rapporto/files/Area13/VQR2004-2010_Area13_RapportoFinale.pdf.

[18] Valutazione della Qualità della Ricerca 2004-2010 (VQR 2004-2010) Comunicato del GEV13 del 2 aprile 2012. http://www.anvur.org/attachments/article/92/gev13_allegati.zip [English translation by the authors].

[19] "Please note that these classifications should not in any way be considered exhaustive of the disciplines of interest of the Area 9, but just want to provide a simple aid to the authors, reducing the effort during product selection. It is possible to submit for evaluation … articles published in journals classified in WoS and not included in the published lists. [In these cases] The classification of the journal will be carried out *ex post*". http://www.anvur.org/attachments/article/87/gev09_criteri.pdf

[20] Because, according to the panel, these journals published only reviews. As a consequence the distribution of journals according to IF changed with respect to the simple distribution reachable through JCR.

[21] This is the description of the use of the citation thresholds in Area 1: "One way to get a rough idea of the thresholds is to go on the ISI Web of Science … site and do an advanced search by specifying year of publication and Subject Area (for example, "Mathematics"). Then refine your search by selecting a Subject Category (for example, still "Mathematics"), sort the list of results based on the number of citations, and you will find the number of citations that allows it to be, respectively, in the first 20%, in the second 20%, and in the upper half of this ordered list. An example of the calculated threshold values in this way for the year 2009 and the Subject Category "Mathematics applied" is in ANVUR document http://www.anvur.org/sites/anvur-miur/files/la_bibliometria_della_vqr.pdf. Remember, however, that the threshold values that will be used in VQR may be different from those so determined, because they will be calculated based on citations received 31 December 2011". FAQ n. 12 in http://www.anvur.org/attachments/article/77/gev01_faq.pdf. [Translation by the authors].

[22] http://www.anvur.org/rapporto/files/Appendici/VQR2004-2010_AppendiceB.pdf, p. 5 [Translation by the authors].

These details are contained in an Appendix of each Area report. These appendices for Areas 1 to 9 were written by inserting specific results in a pre-defined common framework. The procedure for synthesizing the two reviewers reports was therefore described in a nearly identical form. P1 and P2 were identified, respectively, as the numerical scores assigned to an article by a first and a second reviewer; P indicated the "Synthetic evaluation of the scores of the first and second reviewers".[23] In each area report, the procedure for arriving to the synthetic evaluation P was described as follows:

> The reviewers scores [P1 and P2] were then synthesized on the basis of a specific algorithm for Area panel [number], according to which, respectively Excellent products have a score of [numerical score interval]; Good products have a score of [numerical score interval]; Acceptable products have a score of [numerical score interval]; Limited products have a score of [numerical score interval].[24]

For Area 13 the procedure was completely changed. If the opinions of the two referees coincided, the final evaluation was probably[25] automatically defined. If the opinion of the two referees diverged, a complex process started:

> 'The opinion [sic] of the external referees [P1 and P2] was then summarized by the internal Consensus Group: in case of disagreement between P1 and P2, the P index is not simply the average of P1 and P2, but also reflects the opinion of two (and occasionally three) members of the GEV13 (as described in detail in the documents devoted to the peer review process).[26]

In the Area 13 report P was significantly renamed as "evaluation of the Consensus Group".[27] A Consensus group was formed by the two members of the panel in charge to choose reviewers. The work of the consensus groups for reaching P, that is the "evaluation of the Consensus group ", was described as follows:

> The Consensus Groups will give an overall evaluation of the research product by using the informed peer review method, by considering the evaluation of the two external referees, the available indicators for quality and relevance of the research product, and the Consensus Group competences.[28] (ANVUR 2013).

The consensus groups in some cases evaluated also the competences of the two referees, and gave ''more importance to the most expert referee in the research.[29]

---

[23] In Italian: "Valutazione di sintesi dei giudizi del primo e secondo revisore".

[24] Appendices to the Final reports of Area 1-9 are available here: http://www.anvur.org/rapporto/ under the head "Rapporti di Area".

[25] In their comment Bertocchi et al. wrote that Consensus groups might have "effectively graded" the papers" when "the Consensus group disagreed on the arithmetic average of the score", without defining if this disagreement might happen also for two reviewers who agreed on the final merit class of an article but by giving different scores.

[26] http://www.anvur.org/rapporto/files/Area13/VQR2004-2010_Area13_Appendici.pdf, p. 52.

[27] http://www.anvur.org/rapporto/files/Area13/VQR2004-2010_Area13_Appendici.pdf, p. 52. See also p. 459 of (Bertocchi et al. 2015) where P is called "final evaluation of the Consensus group".

[28] http://www.anvur.org/rapporto/files/Area13/VQR2004-2010_Area13_Appendici.pdf, p. 65 [Translation by the authors].

[29] http://www.anvur.org/rapporto/files/Area13/VQR2004-2010_Area13_RapportoFinale.pdf, p. 15, This description of the procedure was already reported in the footnote n. 26 of (Alberto Baccini and De Nicolao 2016).

**References**


Baccini, A. (2016). Napoléon et l'évaluation bibliométrique de la recherche. Considérations sur la réforme de l'université et sur l'action de l'agence national d'évaluation en Italie. *Canadian Journal of Information and Library Science-Revue Canadienne des Sciences de l'Information et de Bibliotheconomie, 40*(1), 37-57, doi:10.1353/ils.2016.0003.

Baccini, A., & De Nicolao, G. (2016). Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise. [journal article]. *Scientometrics*, 1-21, doi:10.1007/s11192-016-1929-y.

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C. A., & Peracchi, F. (2015). Bibliometric evaluation vs. informed peer review: Evidence from Italy. *Research Policy, 44*(2), 451-466, doi:10.1016/j.respol.2014.08.004.

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C. A., & Peracchi, F. (2016). Comment to: Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise. [journal article]. *Scientometrics*, 1-5, doi:10.1007/s11192-016-1965-7.