Rejoinder to: “Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise”, forthcoming in Scientometrics.

The paper by Baccini and De Nicolao (BD), in particular its fifth section – entitled “A self-fulfilling experiment” –, is largely a comment to a paper by Bertocchi, Gambardella, Jappelli, Nappi and Peracchi (2015) recently published in Research Policy (RP).[1] The authors of the RP paper were directly involved in the activities of the Italian research evaluation 2004–10 (VQR) for the area “Economics and Statistics”. Many of the points raised by BD were already addressed in the RP paper. Other points are either incorrect or not supported by evidence. Below is a short summary of the key issues.

The fifth section in BD attempts to explain why the results of Area 13 (Economics and Statistics) are different from those of other scientific areas, in particular why the agreement between bibliometric evaluation and peer review is higher in Area 13. As explained in detailed in the RP paper, Area 13 is in many ways special. In the introduction of the RP paper we write: “The area that we consider is particularly interesting because, at least in Italy, it lies in between the “hard” sciences on the one hand and the humanities and social sciences on the other hand. While in the former most research is disseminated through academic journals and is therefore covered by bibliometric databases, the latter are characterized by a more fragmented literature and more frequent publishing in books and other outlets (Hicks, 1999), so that bibliometric databases are incomplete or almost entirely missing.”

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. More specifically, we list some of the differences in the evaluation process between Area 13 and other areas, that are discussed in more detail in the RP paper:

[1] The second section of BD summarizes the rules of the VQR. A similar description is already available for an international readership in Sections 2 and 3 of Ancaiani et al. (2015). In the third and fourth sections BD claim that the value of kappa for Area 13 (0.54) is statistically different (and higher) from other areas. This statistical difference is already apparent from Table 5 in Cicero et al. (2013), which displays confidence intervals for kappa. Cicero et al. (2013) and Ancaiani et al. (2015) report additional results – that BD have chosen to ignore – based on a test for systematic bias between bibliometric analysis and peer review.

- 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.

- The panel evaluating Area 13 based the classification of journals on a combination of five-year Impact Factor (5IF), five-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.

- 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).

- Area 13 included in the journal list also journals not included in WoS, by an imputation method described in the RP paper.

The reasons why the panel for Area 13 decided to follow a somewhat different approach stems naturally from what we stated earlier, namely, that this area presents notable peculiarities compared to other areas – particularly, the heterogeneity of publications, methods and styles of the different sub-fields that called for specific approaches to evaluation. Most importantly, all these changes have been widely discussed and approved within the Area 13 panel, consisting of 36 members. It is also worth stressing again that the purpose of the Italian evaluation exercise was not to make comparisons across area. As a result, not only did the Area 13 panel feel that it did not have to align to the criteria of other areas, but a good deal of the discussion has been, in fact, on how to adopt different approaches from other areas to properly take into account the peculiarities of Area 13.

In short, the evaluations for Area 13 cannot be directly compared to the evaluations for other areas, and differences between kappa values might arise for many reasons which, given the different experiments in the different areas, cannot be explored with the data at hand. Furthermore, the RP paper compared bibliometric analysis and peer review also by statistical analysis of the bias. This piece of evidence is completely ignored by BD.

As already noted, the purpose of the VQR was <u>not</u> to produce evaluations that could be compared across areas. This is very clear from the VQR Call, which states that an evaluation of different universities, research centers and departments was possible only within each area. Of course, the result of the comparison between bibliometric analysis and peer review in Area

13 is debatable. And in fact in the RP paper we are very cautions in interpreting the evidence.

BD are well aware of the many differences between Area 13 and other areas, but conclude that the comparison between bibliometric analysis and peer review in Area 13 is “fatally flawed” on the basis of five claims. Identical arguments were previously published by Baccini alone in an Italian journal (Baccini 2014 in Statistica & Società). Some of these claims, which we list below, are either incorrect or not based on any evidence; others are already discussed in our RP paper.

1. The first issue raised by BD is that the sample of journal articles is not a random sample of the journal articles submitted to the VQR, because some researchers could request that the paper is peer reviewed. However, as explained in the RP paper, the sample is a random sample of papers published in the journals included in the list ranked by the panel. Indeed, all papers published in the journals ranked by the panel were evaluated by bibliometric analysis. 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.

2. The second claim is that panel members and reviewers knew that a journal article sent in peer review was part of the experiment. The list of journals in Area 13 included about 2,000 journals, but the list was far from complete. For instance, it did not include recent journals for which 5IF and AIS were not available, journals not traceable on the web, etc. As explained in the final report of Area 13, the panel received 6,816 journal articles for evaluation. Of these, 5,681 articles were evaluated by bibliometric analysis and 1,135 by peer review. In addition, out of the 5,681 articles, 590 were evaluated also by peer review (those part of the “experiment”). So, in total, 1,135+590=1,725 journal articles were evaluated by peer review. Furthermore, the final evaluation of a journal article depended also on the number of citations received, which ex ante was unknown not only to referees, but also to panel members.

3. The third claim is that the experiment did not compare anonymous manuscripts in peer review with bibliometric indicators, and that referees knew the ranking of journals. This is fully acknowledged in our RP paper: “First and foremost, as noted repeatedly in this paper, 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. As a

result, we can say nothing about the correlation between the primary factors that link them. The goal of the VQR exercise is to evaluate published work of Italian academics between 2004 and 2010, and therefore this limitation is imposed on us by the very structure and goals of the VQR exercise. Future VQR waves could compare bibliometric analysis and peer review using anonymized published material so that neither the publication outlet nor the name of the authors are revealed to the reviewers. The researchers could then identify the correlation produced by the two independent evaluations, after removing the impact of the information about the publication source. To be sure, even if properly anonymized, this exercise will not be straightforward, as reviewers can figure out the identity of the published paper and authors. Moreover, anonymizing papers requires pre-publication texts that may only be provided by the authors or the journals. Second, it is difficult to generalize our results to other disciplines."

4. The fourth claim of BD is that the panel evaluated internally the majority of the papers of the experiment. The claim of BD is: "at least 326 articles out 590 (55.3%) considered for the experiment were evaluated not by referees, but by the consensus groups". This is incorrect. As in all other research areas, there were clear rules binding the consensus groups (one consensus group was formed for each paper evaluated by peer review), which were actually enforced through the web platform common to all areas. Each consensus group was bound by the two reports and by default used the arithmetic average of the scores assigned by the two referees. In case of disagreement between the two referees, the final score was bound to be between the minimum and the maximum evaluation of the referees. For instance, for a paper with two reports – scoring A and B - the final evaluations of the consensus group was bound to be either A or B (C and D were not allowed). One could argue that there were cases in which the consensus group effectively graded the papers. This occurred when (i) the two reports were so different that one referee assigned the minimum score (D) and the other the maximum score (A), and (ii) the consensus group disagreed on the arithmetic average of the score (the default solution). This is the typical situation faced by a journal editor when one referee is very happy about a paper, and the other referee suggests a rejection. Out of the 590 papers of the experiment, we had 15 such cases (RP, Table 11). So one could argue that <u>at most</u> 15 papers (not 326) were evaluated by the panel itself, as in these cases the two reports were not informative.

5. In their final claim, BD hint that panel members coordinated with referees to increase the agreement between bibliometric evaluation

and peer review. This point therefore refers to the integrity of the panel. Let us just note that it would be impossible to induce 36 high-profile researchers, dispersed in 20 different institutions and 3 continents, and over 700 referees (of which 50% active in institutions outside Italy, scattered around the world), to manipulate the results of the experiment.

We are also well aware of the specificities of the various disciplines within Area 13 mentioned in the last paragraph of the fifth section of the BD paper. In the RP paper we write: “Even within our sub-areas, we found important differences between the two approaches to research evaluation. These differences could arise for a number of reasons. First, there may be differences in refereeing style across subject areas: for example, referees may be less generous in some areas than in others. Second, the reliability of journal ranking may differ across areas: for example, the ranking of journals may be more generous (e.g., placing more journals in the top class) in Economics and Management relative to other sub-areas. Finally, the available sample size may limit the power of statistical tests, as in the case of History. Future research could focus on the analysis of these differences and possibly control better for heterogeneity using larger sample size.”

To conclude, the claim of BD (the experiment is “fatally flawed”) is not supported by evidence. On the contrary, the experiment was performed with clearly documented methodology and data. Of course, as in any experiment, one can disagree on the design and interpretation of the results, and propose that the experiment be performed in different ways. And of course one can claim that different approaches or methods may lead to different results, and argue in favour of these alternative approaches. This would not point to a “flaw” of the experiment, however, but simply to a different experiment. On top of all this, as we have tried to explain with this note, even if we look forward to new experiments suggesting a better research design of our problem, we think that the focus of BD on the differences between our results and those of other areas are obvious conclusions given the obvious differences in the approach of Area 13, which is amply documented in the RP paper and deliberately pursued by the panel to reflect the peculiarity of this area.

We understand that our reply may give rise to some further reply by the authors. Not only do we understand, but we also wish that our research design (like any other) be subject to the deep scrutiny by the academic community. However, we feel that sometimes the tone used by BD – and in particular the allegation that our panel has manipulated the data (see

point 5 above) – has gone beyond the threshold of the mutual professional respect that should guide scientific debate.

Graziella Bertocchi, Università di Modena and Reggio Emilia
Alfonso Gambardella, Università Bocconi
Tullio Jappelli, Università di Napoli Federico II
Carmela Anna Nappi, ANVUR
Franco Peracchi, Università di Roma Tor Vergata

References

Ancaiani, A., Anfossi, A. F., Barbara, A., Benedetto, S., Blasi, B., Carletti, V., Cicero, T., Ciolfi, A., Costa, F., Colizza, G., Costantini, M., Di Cristina, F., Ferrara, A., Lacatena, R. M., Malgarini, M., Mazzotta, I., Nappi, C. A., Romagnosi, S., Sileoni, S. (2015). Evaluating scientific research in Italy: The 2004-10 research evaluation exercise. Research Evaluation 24(3), 242-255.

Baccini, A. (2014). La VQR di Area 13: una riflessione di sintesi. Statistica & Società 3(3), 32–37.

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.

Cicero, T., Malgarini, M., Nappi, C. A., Peracchi, F. (2013). Bibliometric and peer review methods for research evaluation: A methodological appraisement (in Italian). MPRA (Munich Personal REPEC Archive).