Baccini, A., and De Nicolao, G. (2021) Just an artifact? The concordance between peer review and bibliometrics in economics and statistics in the Italian research assessment exercise. Quantitative Science Studies. Advance Publication. https://doi.org/10.1162/qss_a_00172

Just an artifact? The concordance between peer review and bibliometrics in economics and statistics in the Italian research assessment exercise

Alberto Baccini[1]* Giuseppe De Nicolao[2]†

[1]Department of Economics and Statistics, University of Siena, Italy
[2]Department of Electrical, Computer and Biomedical Engineering,University of Pavia, Italy

Abstract

During the Italian research assessment exercise (2004-2010), the governmental agency (ANVUR) in charge of its realization performed an experiment on the concordance between peer review and bibliometrics at an individual article level. The computed concordances were at most weak for science, technology, engineering and mathematics. The only exception was the moderate concordance found for the area of economics and statistics. In this paper, the disclosed raw data of the experiment are used to shed light on the anomalous results obtained for economics and statistics. In particular, the data permit to document that the protocol of the experiment adopted for economics and statistics was different from the one used in the other areas. Indeed, in economics and statistics a same group of scholars developed the bibliometric ranking of journals for evaluating articles, managed peer reviews and formed the consensus groups for deciding the final scores of articles after having received the referee's reports. This paper shows that the highest level of concordance in economics and statistics was an artifact mainly due to the role played by consensus groups in boosting the agreement between bibliometrics and peer review.

Keywords: bibliometrics, peer review, research assessment exercise, replication study, economics and statistics, Italy

Alberto Baccini and **Giuseppe De Nicolao**: Conceptualization, Methodology, Formal Analysis, Writing - Original draft, Writing - Review & editing.

*alberto.baccini@unisi.it. ORCID: 0000-0003-0293-482X. Alberto Baccini is the recipient of grants by the Institute For New Economic Thinking Grant Institute For New Economic Thinking Grant ID INO17-00015 and INO19-00023.

†ORCID: 0000-0002-3712-9911

Peer Review: https://publons.com/publons/10.1162/qss_a_00172

1 Introduction

During the research assessment exercise for the years 2004-2010, the Italian governmental Agency for Evaluation of Universities and Research (ANVUR) performed an experiment on the agreement between peer review and bibliometrics at an individual article level (for a recent review of literature see (Baccini, Barabesi, & De Nicolao, 2020)). The experiment involved all the fields of science, technology, engineering and mathematics, plus economics and statistics. The design of the experiment was apparently very linear: a stratified random sample of about 10 thousands journal articles was evaluated by applying bibliometric indicators and by peer review; the degree of agreement between the scores obtained with the two systems of evaluation was then estimated by using weighted Cohen's kappa. The overall results of the experiment were published not only in the official reports (ANVUR, 2013), but also as journal articles authored by researchers affiliated to ANVUR or appointed to carry out the experiment (Ancaiani et al., 2015). For the field of economics and statistics, results of the experiment and a big part of the official report were published by *Research Policy* as a research paper authored by some of the members of the panel appointed by ANVUR to carry out the research assessment in the field (Bertocchi, Gambardella, Jappelli, Nappi, & Peracchi, 2015).[1]

In a nutshell, results of the experiment were generally presented in the official reports as successful by stating that there is a "more than adequate concordance" between bibliometrics and peer review (ANVUR, 2013). This "fundamental agreement" (Ancaiani et al., 2015) would support the use of the so called "dual system of evaluation" adopted in the research assessment, consisting in the interchangeable use of bibliometrics and peer review for evaluating journal articles. Economics and statistics presented the highest level of agreement between bibliometrics and peer review. Results of the Italian experiment are cited as a solid evidence of the agreement between bibliometrics and peer review at an individual article level in scientometric literature and in the discussion about reliability of research assessment (among the others Mittermaier (2020); Thomas, Nedeva, Tirado, and Jacob (2020); Fassin (2021); Rousseau and Rousseau (2021)).

Doubts about the reliability of the whole Italian experiment and especially for the field of economics and statistics were raised by Baccini and De Nicolao (2016a, 2017a, 2017b, 2016b) on the basis of the official published data. In a first paper (Baccini & De Nicolao, 2016a), they highlighted an anomalous high level of agreement reached for economics and statistics with respect to all the other research areas. They argued that it was due to substantial modifications of the protocol of the experiment in this field with respect to the other areas. They described these modifications on the basis of ANVUR official documents. Bertocchi et al. (2016) denied the existence of the modifications. Baccini and De Nicolao replied by confirming all their claims, but they were limited by the

[1] Both in the case of the overall results and of economics and statistics, no indications are available that permit to distinguish between the official positions of ANVUR and the views expressed by the authors of the published articles.

impossibility to verify some conjectures on the basis of the raw data (Baccini & De Nicolao, 2016b). Afterwards, Baccini and De Nicolao (2017b) documented statistical problems in the experiment and factual errors in the way in which it was reported by Ancaiani et al. (2015). They replied by correcting some errors in their paper and by denying the relevance of the statistical problems (Benedetto, Cicero, Malgarini, & Nappi, 2017). All of these issues could have been easily resolved if ANVUR or the authors of the papers had disclosed the raw data of the experiment.[2]

In March 2019 only, ANVUR decided to disclose the raw data of the experiment. [3] This disclosure has permitted Baccini et al. (2020) to reconsider in full the experiment by providing the correct design-based setting for it. They showed that "for each research areas of science, technology, engineering and mathematics the degree of agreement between bibliometrics and peer review is – at most – weak at an individual article level". They confirmed also the anomalous high value of agreement for the area of economics and statistics.

On the basis of the raw data now available, this paper aims to finally establish: (i) if the protocol of the experiment adopted for economics and statistics was different from the one adopted in the other areas; (ii) if this difference was responsible of the anomalous agreement in economics and statistics; (iii) if the description of the experiment published in Bertocchi et al. (2015) is correct; (iv) if the claims about the experiment contained in Bertocchi et al. (2016) are true of false. In section 2 of the paper a short description of the experiment is provided. Section 3 illustrates the interventions of the so-called consensus groups for scoring the articles of the experiment. In sections 4 the effect on the agreement between peer review and biblometrics of the different protocol adopted in Area 13 is estimated. The final section 5 discusses the results and the blackgeneral lessons that can be drawn for research assessment and research policy.

2 A short description of the protocol of the experiment

The ANVUR experiment involved 10 research areas of science, technology, engineering and mathematics, plus economics and statistics. For each area a panel managed the evaluation. Each panel was composed by a number of scholars proportional to the dimension of the area. For each area, ANVUR selected a random sample of journal articles. These articles were scored both by bibliometrics and by peer review. There were four possible letter scores: A (Excellent), B (Good), C (Acceptable) and D (Limited).

[2]The authors of this paper requested the data to the President of ANVUR, at the date prof. Stefano Fantoni (mail sent on Febrary 10th 2014). They never received a reply.

[3]The mail from one of the author to Prof. Paolo Miccoli, president of ANVUR, containing the request is dated from March 12th 2019. The decision of disclosing the data was communicated by mail dated March 26th 2019; the access to the data was open April 9th 2019. Data can be downloaded from https://doi.org/10.5281/zenodo.3727460.

For all the areas, except economics and statistics (Area 13), the bibliometric scores were attributed accordingly to an algorithm. It combined the number of citations of an article and a bibliometric indicator of the impact of the journal where it was published. If the two indicators were coherent (e.g. high number of citations and high impact factor) the articles received a score. If the two were incoherent (e.g. high number of citation and low impact factor) the algorithm returned an inconclusive score "IR". While in the research assessement IR papers were scored by peer review, in the experiment they were simply dropped from the sample (for a discussion of the statistical problems induced by this procedure see Baccini et al. (2020)).

For Area 13 only, the bibliometric algorithm consisted in scoring a paper on the basis of the journals where it was published. To this end, the Area 13 panel directly developed a ranking of journals organized in four classes from *A* to *D*.[4] As a consequence, differently from the other areas, in Area 13 there were no articles with inconclusive bibliometric score and no papers were dropped from the experiment. Columns 1 and 2 of Table 1 report, stratified by research areas, respectively the size of the experiment sample and the size of the subsample after the removal of the IR papers.

As for the peer review, each article was assigned to two of the members of the area panel. They formed a so-called Consensus Group (CG). In turn, each of the two members of the CG selected a referee who evaluated the article by assigning a numerical score according to a pre-defined format. The format required that the referee evaluated a paper according to three criteria: relevance; originality/innovation; internationalisation. Each criterion received a partial score; the sum of the three scores represented the final score assigned by a referee to a paper. The two referee's reports were indicated as "P1" and "P2". In the Areas 2, 3, 6, 7, 8a and 13, referees were required to score each criterion in a scale from 1 to 9 points; hence the total score assigned by a referee to a paper ranged from 3 to 27 points. In the Areas 1, 4, 5, 9, referees were required to score each criterion in a scale from 0 to 3 points; hence the total score assigned by a referee to a paper ranged from 0 to 9 points. CGs received the two referee's reports P1 and P2 and "synthesized [them] in a final evaluation" (P) (See ANVUR (2013): Appendix B, p. 5). For all the areas, except for the Area 13, this final evaluation was based on "algorithms specifically defined by each Area panel" (See ANVUR (2013): Appendix B, p. 5). It appears that the final evaluation P was simply the average of the two numerical scores P1 and P2 assigned by the two referees. This average was then converted in one of the four final scores P, according to the two "conversion grids" reported in Table 2.[5]

[4]The methodology adopted for the classification is available in (Bertocchi et al., 2015). An Italian administrative court conclusively invalidated the procedure and methodology adopted for the journal ranking, because of "failure to carry out an investigation, misinterpretation of facts and failure to state reasons" (Tribunale Amministrativo del Lazio, 30/10/2017, n. 10805/2007; https://tinyurl.com/y6sqwo4p).

[5]The descriptions of the procedures adopted by each Area panel are in ANVUR (2013) (see the Appendice A of each area report).

Table 1: Sample and subsample size, number of articles with a final scores coincident with the scored agreed by two referees (P=P1=P2) number of articles for which two referees indicated non concordant scores (P1≠P2), number of articles for which the consensus groups changed two concordant referees' scores P≠P1=P2, total number and share of articles scored after an intervention of a consensus group.

Scientific Areas	Sample (1)	Subsample (2)	P=P1=P2 (3)	P1≠P2 (4)	P≠P1=P2 (5)	Scored by CG (6=4+5)	Scored by CG (%) (7=6:2)
Area 1 - Mathematics and Informatics	631	438	207	230	1	231	52.7
Area 2 - Physics	1412	1212	513	696	3	699	57.7
Area 3 - Chemistry	927	778	339	438	1	439	56.4
Area 4 - Earth Sciences	458	377	149	228	0	228	60.6
Area 5 - Biology	1310	1058	433	623	2	625	59.1
Area 6 - Medicine	1984	1603	607	994	2	996	62.1
Area 7 - Agricultural and Veterinary Sciences	532	425	134	290	1	291	68.5
Area 8a - Civil Engineering	225	198	85	112	1	113	57.1
Area 9 - Industrial and Information Engineering	1130	919	378	540	1	541	58.9
Area 13 - Economics and Statistics	590	590	255	326	9	335	56.8
Areas 1-9	8609	7008	2845	4151	12	4163	59.4
All areas	9199	7598	3100	4477	21	4498	59.2

Source: elaboration on ANVUR data.

Table 2: The conversion grids adopted for transforming the numerical scores in the final letter score P. Numerical scores are computed by averaging the scores P1 and P2 resulting from peer review. The range of numerical scores are indicated as intervals.

	Areas 2, 3, 6, 7, 8a and 13	Areas 1, 4, 5 and 9
P	Score range	Score range
A	[8-9]	[23-27]
B	[6-8[	[18-23[
C	[5-6[	[15-18[
D	[0-5[	[3-15[

Area 13 also adopted a conversion grid (see the note 24 of Bertocchi et al. (2015)), but, at the same time, it adopted also a more elaborated protocol for CGs' decisions, by permitting a more flexible treatment of the referees' reports. This protocol is described in the official report as follows:

> The opinion [sic] of the external referees 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 (See ANVUR (2013): Area 13 Report, Appendix A, p. 52).

The point was stressed in more than one part of the official reports:

> 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 (ANVUR, 2013).

And again:

> 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 field" (see ANVUR (2013) Area Report, p. 15 translation from Italian by the authors).

According to Baccini and De Nicolao (2016a), the main difference between the protocol of the experiment for Area 13 and for the other areas consisted properly in letting that consensus group might consider so many elements for the final decisions.

Moreover, also the information available to the members of the CGs were different in Area 13 with respect to the other areas: (i) the member of the CGs in Area 13 knew that the journal articles for which they had to arrange a peer review were the ones selected for the experiment. Indeed, all the articles submitted to the research assessment and published in journals listed in the

ranking developed by the area panel received an automatic score. This was not the case in the other areas, where panels had to arrange peer reviews not only for the articles of the experiment, but also for the ones submitted to the research assessment and classified as IR (inconclusive rating) by the bibliometric algorithm; (ii) the CG members knew the final bibliometric score of articles, while in the other areas the CGs might knew only the bibliometric indicators informing the bibliometric algorithm. The information about the bibliometric score of each article might have been used by the CGs when they chose the referees and when they decided the final peer review score of each article.

In addition, there were also differences regarding information available for the referees. The referees of Area 13 were possibly aware that they were participating to the experiment for the same reason discussed above for the members of the panel. Indeed, in Area 13 all journal articles submitted to the research assessment were automatically scored according to the journal rank. So if a referee received a journal article for evaluation, it was obviously one of the sample extracted for the experiment. In the other areas, as anticipated, referees received many journal articles because they had an inconclusive bibliometric rating. Hence, it was impossible for referees in the other areas to know if an article was part of the sample of the experiment. Differently again from the other areas, the referees of Area 13 knew also the bibliometric classification of the articles, indeed:

> The referees were provided with the panel journal classification list and the actual or imputed values of IF, IF5 [five years impact factor] and AIS [Article influence score] (Bertocchi et al., 2015).

By having access to the ANVUR raw data, it is possible to verify in details how these modifications of the protocol impacted on the experiment conducted in Area 13. In particular, it is possible to verify if and how the more active role of the consensus groups impacted on the results of the experiment with respect to the other areas.

3 The role of consensus groups: how many papers have they evaluated?

The first question is how many papers required an intervention of the CGs. For answering, the total number of papers can be partitioned in three non-overlapping subsets. The three sets are reported in Table 1 stratified by scientific areas. The three sets are composed by:

1. papers for which two referees indicated a concordant score that was also confirmed as the final score (Table 1, column 3: P=P1=P2);

2. papers for which two referees indicated discordant scores (Table 1, column 4: P1≠P2);

3. papers for which the final score was different from the one agreed by the two referees (Table 1, column 5: P≠P1=P2).

As previously noted, when the two referee's reports did not coincide, the final peer review score of an article required an intervention of the CGs. Then the total number of articles for which the final score was obtained after a CG intervention can be obtained by summing up the columns 4 and 5 of Table 1: the sum is reported as the column 6 of Table 1. The expression "Scored by CG" used in this paper simply indicates that the final score was decided after a CG intervention. This intervention might have consisted in confirming the average between P1 and P2, as calculated by the algorithm; or in deciding the final letter score by modifying the scores indicated by the referees.

On the whole experiment, the share of papers finally requiring a CG intervention was 59.2% of the sample. In Area 13 this share was 56.8%, only a bit lower than the average. From this point of view and on the whole, in Area 13, CGs did not intervene more actively than in the other areas. Nonetheless, Area 13 shows the highest number of articles for which CGs changed a concordant score of the two referees. In Area 13 CGs changed a concordant score of the two referees for 9 articles out of 590, representing the 1,52% of the total articles in the area sample. In all the other areas, CGs changed just 16 articles out of 7,598, representing the 0,16% of the total experiment sample. This may be considered as a first clue of an attitude of the Area 13 panel to intervene in scoring papers more actively than the panels of the other areas.

Table 1 finally shows that Baccini and De Nicolao (2016a) even underestimated the number of 326 papers scored after a CG intervention in Area 13. Bertocchi et al. (2016) had contested this estimate by stating that:

> one could argue that at most 15 papers (not 326) were evaluated by panels itself.

How is it possible to have this very big misalignment on a basic fact? Bertocchi et al. limited their attention to 15 articles for which they argued that the CGs

> effectively graded the paper. 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 CGs disagreed on the arithmetic average of the score (the default solution).(Bertocchi et al., 2016)

This very restrictive claim about the intervention of the CGs is at odds with official reports and Bertocchi et. al.'s own description of the experiment (Bertocchi et al., 2015). Table 1 finally shows that it is also strictly falsified by the data, since in addition to the 15 articles for which the referees were in maximum disagreement, as we have seen, CGs "directly graded" 9 other papers for which the two referees indicated a concordant score.[6]

[6]In particular: 4 papers, concordantly scored B by two referees, were classified as A by the

It remains to clarify how invasive the interventions of the consensus groups were in defining the final score P. The most invasive CG's intervention consisted obviously in changing a score agreed by two reviewers. But CGs might have adopted a less invasive strategy by assigning a final P score without applying rigidly the "conversion grid" reported in Table 2. If ANVUR had disclosed the numerical scores of the referees' reports instead of P1 and P2 only, it would be possible to trace precisely the intervention of the CGs by comparing the score P with the average of the numerical scores attributed by two referees. In every case, disclosed data permit to show that CGs, especially in the Area 13, graded some articles out of the rules as defined in the official reports.

Indeed, it is possible to roughly define for the final scores P, lower and upper bounds within which CG interventions respect the rules of the assessment, by considering the conversion grid adopted in each area. Lower bounds for CG decisions are computed as follows. The first step consisted in calculating the minimum average associated with each possible combination of P1 and P2. Consider, for example, that a first referee assigns a score P1=A and a second referee a score P2=B. The minimum average is calculated under the hypothesis that both reviewers assigned the minimum numerical score to the paper: the first referee judged the paper as A by assigning a numerical score of 23; while Referee 2 judged the paper as B by assigning the minimum numerical score of 18. Therefore the minimum possible average for a paper judged A by one referee and B by the other is (23+18)/2= 20.5, corresponding in the conversion grid to a final score B. Hence, for respecting the rules of the assessment, the final score P of a paper receiving P1=A and P2=B should be A or B. Upper bounds are analogously computed. The maximum average can be calculated by considering that both referee assign the maximum scores for A and B, respectively 27 and a bit less than 23. Therefore the maximum possible average for a paper judged A by one referee and B by the other is (27+23)/2= 25, corresponding in the conversion grid to the letter score A. This is the upper bound for CG decision. Table A1 in the Supplementary material reports the upper and lower bounds computed for the two conversion grids adopted in different areas. Note that in the case of P1=A and P2=C the final letter score P is necessarily B since the minimum possible average is (23+15)/2=19 and the maximum possible average is (27+18)/2=22.5, and both numerical scores correspond to the letter score B; analogously for P1=A and P2=D the upper bound is P=B, since the maximum average score is (27+15)/2=21 corresponding to a letter score B.

Table 2 reports the number of papers scored by CGs out of the bounds of Table A1, i.e. the number of papers scored not respecting the declared rules of the assessment. The consensus groups of the Areas 1-9 did not respect the bounds for a total of 28 out of 4163 papers (0.67%). In Area 13 the consensus groups did not respect the bounds for 23 out of 335 papers (6.87%). The big part of these papers received a final score P=A. This is a second clue indicating that in Area 13 consensus groups have a more active attitude by deciding the

CGs; 2 papers respectively scored C and D by two concordant referees, were finally classified as B by the CGs, and 3 papers concordantly scored D by referees were classified C by the CGs.

Table 3: Number of papers scored by consensus groups out of the bounds of the research assessment, as reported in the Table A1 of Supplementary material. The percentage is calculated over the total number of papers requiring the intervention of the consensus groups (Table 1, column labelled: "Scored by CG").

Final P score	A	B	C	D	Total	%
Area 1 - Mathematics and Informatics	0	0	1	0	1	0.43
Area 2 - Physics	0	0	3	3	6	0.86
Area 3 - Chemistry	0	1	0	0	1	0.23
Area 4 - Earth Sciences	0	0	0	1	1	0.44
Area 5 - Biology	0	1	3	0	4	0.64
Area 6 - Medicine	6	2	2	1	11	1.10
Area 7 - Agricultural and Veterinary Sciences	2	0	0	0	2	0.69
Area 8a - Civil Engineering	0	1	0	0	1	0.88
Area 9 - Industrial and Information Engineering	0	1	0	0	1	0.18
Area 13 - Economics and Statistics	16	3	4	0	23	6.87
Areas 1-9	8	6	9	5	28	0.67
All areas	24	9	13	5	51	1.13

Source: elaboration on ANVUR data.

final letter score P, with respect to the other areas.

More in general, Figure 1 permits to visually compare the interventions of CGs in deciding the final score of a paper in the different areas of the experiment. The graph is organized in 40 facets representing ten areas (columns) and the four final P-scores (rows). Each panel represents the scores P1 (x-axis) and P2 (y-axis) in a given area for a given final peer review score P. The size of each point indicates the proportion of articles finally scored P in the given area. Blue points indicates articles scored by respecting the declared bounds of the assessment, while red points indicates articles scored not respecting the bounds. Tables A5.1 and A5.2 of the Supplementary material report the data used for building Figure 1.

Consider the top left panel. In Area 1 (mathematics and informatics) most of the articles with a final P-score A in the experiment were concordantly classified as A by the two reviewers P1 and P2; a few of these articles were also scored A by one of the referees and B by the other. No articles scored less than B by one of the referee was finally scored A by the CG. Consider now the top right panel. In Area 13, there were many papers scored less than B by one of the two referees that were finally classified as A by the CGs. It is apparent that CGs scored A some papers for which concordantly the two referees had indicated a score B; and also some papers for which none referee indicated a score A. From the visual inspection of the Figure 1 it is apparent that CGs of Area13 behave differently from the other areas, by adopting a greater flexibility than in the other areas in the conversion of the referees' scores P1 and P2 in the final score

P.

In sum, the consensus groups of Area 13 managed a share of papers similar to all the other areas. Data documented that they had a more active attitude both in modifying the scores agreed by the referees, and in scoring the papers outside the bounds defined in the rules of the research assessment. Moreover, they tended to interpret more flexibly than in the other areas the rules for converting the referees' reports in the final P score.

4 How much of the agreement between peer review and bibliometrics was induced by CG decisions?

On the basis of ANVUR data it is now possible to shed light also on the central question about the experiment: how much of the agreement between peer review and bibliometrics depended on the decisions of the CGs, i.e. how much of the agreement was induced by the scores defined by the members of the panel. From Table A1, it is evident that even while respecting the bounds, CGs had a good margin of flexibility in deciding the final score P. For instance, after having received two discordant peer review reports indicating P1=B and P2=D, a CG can decide a final P score B or C or D, perhaps in accordance with the bibliometric score.

For measuring the role of CGs' decisions in determining the agreement between peer review and bibliometrics, it is possible to build two indicators.

The first indicator, reported in Table 4, is the percentage of CG decisions that produced a final score in agreement with bibliometric score. It is calculated as the ratio between the number of papers scored by CGs in agreement with bibliometrics, and the total number of papers scored by CGs. In Area13, GCs attributed a score in agreement with bibliometrics for 45,4% of the papers scored by CGs (152 papers out of 335). In the other nine areas, this share ranges from a minimum of 17.6% in Area 9 to a maximum of 27.9% in Area 6. In the other nine areas taken togheter, CGs attributed a score in agreement with bibliometrics for only 23.7% of the papers scored by CGs (986 concordant papers out of 4,163). The agreement induced by CG decisions was anomalous with respect to the other areas mainly for the subset of articles scored less than A. This indicates that in Area 13 consensus groups' interventions boosted the agreement between peer review and bibliometrics for the set of papers receiving a final score less than A. In particular, in Area 13 the share of concordant C papers was almost three times as much as in the other areas; and the share of concordant D papers was a bit less than the double with respect to the other areas.

The second indicator, reported in Table 5, is the share of papers for which the agreement between peer review and bibliometrics was due to the decisions of the CG. It is computed as the ratio between between the number of papers scored by CGs in agreement with bibliometrics, and the total number of papers for which there is agreement between peer review and bibliometrics. In Area 13

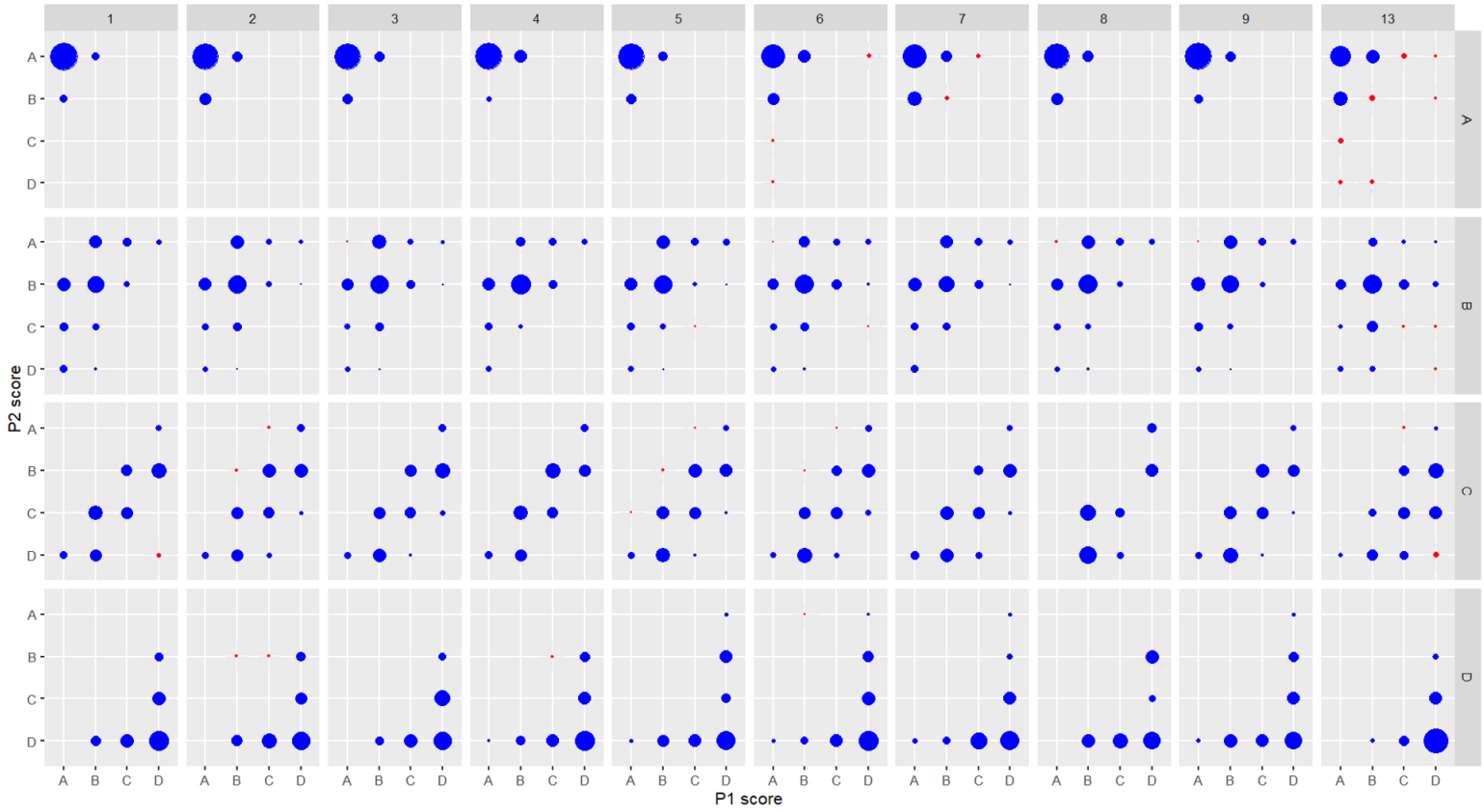

Figure 1: Visual comparison of the interventions of CGs in deciding the final score of a paper in the different areas of the experiment. The graph is faceted according to ten disciplinary areas (columns) and the four final P scores (A, B, C, D) . Each panel represents the scores attributed by P1 (x-axis) and P2 (y-axis) in a given area for a given final peer review score P. The size of each point indicates the proportion of articles finally scored P in the given area. Blue points indicates articles with a final score respecting the bounds of the assessment as reported in Table A1; red points indicates articles for which CGs did not respect the bounds.

Table 4: Percentage of papers scored by consensus groups in agreement with bibliometrics over total papers scored by consensus groups, stratified according to final P scores.

Area	A	B	C	D	TOTAL
Area 1 - Mathematics and Informatics	73.3	14.7	20.0	21.3	20.8
Area 2 - Physics	83.1	19.6	10.3	26.3	24.6
Area 3 - Chemistry	75.6	22.6	5.3	25.6	24.1
Area 4 - Earth Sciences	71.4	18.6	8.3	41.3	23.2
Area 5 - Biology	70.6	19.2	8.1	38.4	24.8
Area 6 - Medicine	69.2	27.2	4.6	44.8	27.8
Area 7 - Agricultural and Veterinary Sciences	86.4	9.2	0.0	62.8	20.6
Area 8a - Civil Engineering	92.3	4.6	4.8	28.6	17.7
Area 9 - Industrial and Information Engineering	89.7	13.4	11.3	6.6	17.6
Area 13 - Economics and Statistics	81.0	29.5	29.4	65.5	45.4
Areas 1-9	78.5	18.8	7.5	36.7	23.7
All areas	78.9	19.3	9.5	38.7	25.3

Source: elaboration on ANVUR data.

Table 5: Percentage of papers scored by consensus groups in agreement with bibliometrics, over total number of papers with concordant peer review and bibliometrics, stratified according to final P scores.

Area13	A	B	C	D	TOTAL
Area 1 - Mathematics and Informatics	9.6	61.3	100.0	38.5	26.8
Area 2 - Physics	23.4	57.1	85.0	54.1	38.9
Area 3 - Chemistry	19.1	56.6	83.3	47.6	35.9
Area 4 - Earth Sciences	16.1	42.1	85.7	53.1	42.4
Area 5 - Biology	16.8	49.5	80.0	49.6	38.9
Area 6 - Medicine	28.8	64.9	93.3	49.2	48.9
Area 7 - Agricultural and Veterinary Sciences	34.5	50.0	0.0	61.4	47.2
Area 8a - Civil Engineering	27.9	60.0	100.0	44.4	34.5
Area 9 - Industrial and Information Engineering	17.2	59.7	80.0	26.7	31.0
Area 13 - Economics and Statistics	52.0	55.4	82.1	32.2	48.9
Areas 1-9	20.8	56.9	85.9	49.7	39.5
All areas	23.2	56.8	84.7	46.7	40.5

Source: elaboration on ANVUR data.

there were a total of 311 papers for which peer review and bibliometrics were in agreement; for 152 of these papers, i.e. for a share of 48.9%, the final peer review score was decided by the CGs. In the other nine areas the share was of 39.5% only (986 papers out of 2,857). The anomaly of Area13 was concentrated in the group of papers scored A: in Area13 CGs directly scored more than half (51 out of 98, that is 52%) of the concordant papers against one fifth (241 out of 1160, that is 20.8%) of the other areas. This second indicator shows that more than half of the papers scored as excellent by both bibliometrics and peer review received the final P score after an intervention of the member of the Area 13 panel.

5 Discussion and conclusion

The results of the experiment performed by ANVUR during the Italian research assessment exercise VQR 2004-2010 have a central role in the ongoing discussion about the agreement between peer review and bibliometrics. Indeed, it is probably the most extensive experiment conduceted so far for verifying the concordance between peer review and bibliometrics. Its results were presented as indicating a "fundamental agreement" between peer review and bibliometrics in science, technology, engineering, mathematics and especially in economics and statistics. Despite the early critics to the reliability of the whole experiment, results are currently cited (Mittermaier, 2020; Thomas et al., 2020; Fassin, 2021; Rousseau & Rousseau, 2021) as indicating a solid evidence of a good agreement between peer review and bibliometrics at an individual article level. Actually, when the results of the experiments were replicated in the correct inferential setting, they showed that for science, technology, engineering and mathematics the degree of agreement between bibliometrics and peer review was at most weak at an individual article level (Baccini et al., 2020). The only exception was economics and statistics, where the agreement was moderate.

This work aimed to finally test whether this anomalous result for economics and statistics was due to a substantial modification of the protocol of the experiment with respect to the one adopted in the other areas, as early suggested by Baccini and De Nicolao (2016a, 2016b, 2017b, 2017a).

The data eventually disclosed by ANVUR reveal that the official report published by ANVUR, the text collated from it and published by *Research Policy*, as well as the "final" description provided in Bertocchi et al. (2016), contain partial or even incorrect descriptions of the protocol of the experiment conducted in Area 13.

In particular, in Area 13, the CGs decided the final score of 335 papers out of a total of 590. These 335 included 326 papers for which the two referees were in disagreement, and 9 papers that CGs scored by modifying the concordant score suggested by two reviewers. Therefore, the raw data directly and conclusively falsify the statement by Bertocchi et al. (2016) that in Area 13 "at most 15 papers" were evaluated by CGs.

The raw data also show that for 6.87% of the papers, the Area 13 CGs did

not respect the upper and lower bounds for scoring articles stated in the official reports and in Bertocchi et al. (2015, 2016). In the other nine Areas, the share of scores not respecting the declared bounds was just 0.67%.

Moreover, the ANVUR data show that CGs played a major role in boosting the agreement reached in the experiment of Area 13. In Area 13, 45,4% of the scores given by the CGs agreed with bibliometrics, against a 23.7% in the other areas. In particular, among the papers with a concordant A-score between peer review and bibliometrics, as much as 52% had been scored by the CGS against a 20.8% of the other areas.

In sum, the disclosed raw data of the experiment document that the moderate agreement between bibliometrics and peer review in economics and statistics was an anomalous result produced by the active intervention of the members of the consensus groups in charge of synthesizing peer review reports (ANVUR, 2017).

This conclusion is corroborated by the results of a second experiment, conducted by ANVUR during the national research assessment VQR 2011-2014. In this second experiment the protocol "excluded the intervention" of the consensus groups in the definition of the final peer review P scores, that were instead computed by an algorithm in all the areas (see ANVUR (2017): Appendix B, p. 8 note 4). The replication of the results of this second experiment in the correct inferential setting showed that "when an identical protocol was adopted for all the areas, the agreement for Area 13 was only slightly larger, but still comparable with the other areas" (Baccini et al., 2020). More specifically, in the second experiment the degree of agreement between bibliometrics and peer review is generally even lower than in the first one, by indicating that the agreement between peer review and bibliometrics at the level of individual articles is at most weak in all the considered research areas (Baccini et al., 2020).

In a nutshell, in Area 13 a group of scholars was called to develop a bibliometric ranking of journals for attributing a bibliometric score to articles published in these journals. This same group of scholars was called also to manage peer reviews for the papers published in these journals. Finally, they formed the consensus groups for deciding the final peer review scores of articles after having received the referee reports. To this end, the consensus groups had not only flexible margins for their decisions, but also the freedom to deviate from the rules of the experiment fixing the bounds for scoring articles. Given all these premises, it is hardly surprising that in economics and statistics the agreement between bibliometric and peer review reached a level not recorded in any other area considered in the Italian experiment.

Actually, the decisions of the panel for economics and statistics finished to confirm and strengthen the bibliometric assessment methods it had developed. Recall that for economics and statistics only the bibliometric score of an article was defined on the basis of the journal ranking developed by the panel. As a consequence, a high level of agreement would indicate that the ranking of journals developed by the panel was a good proxy of the quality of articles as revealed by reviewers in their reports. In particular, the choices of the consensus groups delimited the set of excellent documents. If the experiment shows that

the articles rated as excellent by peer review are those published in journals rated as excellent by the panel, a clean and simple criterion of excellence could finally be established: excellent articles are those and only those hosted by a restricted set of supposedly excellent journals.

On a more general level, the evidence of a good agreement would justify the use of journal ranking instead of peer review for evaluating papers in economics and statistics. Indeed, results of the experiment were used for producing policy advices about research evaluation for an international audience (see for example Bertocchi, Gambardella, Jappelli, Nappi, and Peracchi (2014)). The good agreement and the consequent policy advice are expectedly welcome in economics, a scholarly environment particularly fascinated by journal rankings (Heckman & Moktan, 2020). It is well known that the use of journal rankings tends to reinforce existing hierarchies within disciplines (Corsi, D'Ippoliti, & Zacchia, 2019; Heckman & Moktan, 2020; Stockhammer, Dammerer, & Kapur, 2021) and, at the same time, reduce pluralism of research. A growing literature (Lee, Pham, & Gu, 2013; Corsi et al., 2019; D'Ippoliti, 2021) suggests that the reduction of pluralism in economics cannot be considered just as an unintended consequence of research assessment (Rousseau & Rousseau, 2021).

In summary, in light of the raw data disclosed by ANVUR, the current interpretation of the Italian experiment on peer review and bibliometrics agreement should undergo revision and be realigned with the available evidence. The first Italian experiment showed that peer review and bibliometrics have less than weak agreement at an individual article level for the fields of science, technology, engineering and mathematics (Baccini et al., 2020). Moreover, the higher level of agreement in economics and statistics appears to be simply an artifact of the experiment protocol adopted by the group in charge of evaluating economics and statistics. Hence, the results of the Italian experiment cannot be considered as a solid evidence of a special agreement between peer-review and journal ranking, even for the fields of economics and statistics.

Competing interests: The authors have no competing interests to declare.
Funding information:Alberto Baccini is the recipient of grants by the Institute For New Economic Thinking Grant Institute For New Economic Thinking Grant ID INO17-00015 and INO19-00023.
Data availability: Raw data used in the article can be downloaded from https://doi.org/10.5281/zenodo.3727460.

References

Ancaiani, A., Anfossi, A. F., Barbara, A., Benedetto, S., Blasi, B., Carletti, V., ... Sileoni, S. (2015). Evaluating scientific research in italy: The 2004–10 research evaluation exercise. *Research Evaluation*, *24*(3), 242-255. doi: 10.1093/reseval/rvv008

ANVUR. (2013, June, 30 2013). *Rapporto finale. valutazione della qualità della ricerca 2004-2010* (Tech. Rep.).

ANVUR. (2017, 21 febbraio). *Valutazione della qualità della ricerca 2011-2014. rapporto finale* (Tech. Rep.).

Baccini, A., Barabesi, L., & De Nicolao, G. (2020). On the agreement between bibliometrics and peer review: evidence from the italian research assessment exercises. *PLOS ONE*, *15*(11), e0242520. doi: 10.1371/journal.pone.0242520

Baccini, A., & De Nicolao, G. (2016a). Do they agree? bibliometric evaluation versus informed peer review in the italian research assessment exercise. *Scientometrics*, *108*(3), 1651-1671. doi: 10.1007/s11192-016-1929-y

Baccini, A., & De Nicolao, G. (2016b). Reply to the comment of bertocchi et al. *Scientometrics*, *108*(3), 1675-1684. doi: 10.1007/s11192-016-2055-6

Baccini, A., & De Nicolao, G. (2017a). Errors and secret data in the italian research assessment exercise. a comment to a reply. *RT. A Journal on Research Policy and Evaluation*, *5*(1). doi: 10.13130/2282-5398/8872

Baccini, A., & De Nicolao, G. (2017b). A letter on ancaiani et al. 'evaluating scientific research in italy: the 2004-10 research evaluation exercise'. *Research Evaluation*, *26*(4), 353-357. doi: 10.1093/reseval/rvx013

Benedetto, S., Cicero, T., Malgarini, M., & Nappi, C. (2017). Reply to the letter on ancaiani et al. 'evaluating scientific research in italy: The 2004–10 research evaluation exercise'. *Research Evaluation*, *26*(4), 358-360. doi: 10.1093/reseval/rvx017

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C. A., & Peracchi, F. (2014). Assessing italian research quality: A comparison between bibliometric evaluation and informed peer review. *www.voxeu.org*.

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. *Scientometrics*, 349-353. doi: 10.1007/s11192-016-1965-7

Corsi, M., D'Ippoliti, C., & Zacchia, G. (2019). Diversity of backgrounds and ideas: The case of research evaluation in economics. *Research Policy*, *48*(9), 103820. doi: https://doi.org/10.1016/j.respol.2019.103820

D'Ippoliti, C. (2021). Many-citedness: citations measure more than just scientific quality. *Journal of Economic Surveys*. doi: 10.1111/joes.12416

Fassin, Y. (2021). Does the financial times ft50 journal list select the best management and economics journals? *Scientometrics*, *126*(7), 5911-5943.

doi: 10.1007/s11192-021-03988-x

Heckman, J. J., & Moktan, S. (2020, June). Publishing and promotion in economics: The tyranny of the top five. *Journal of Economic Literature*, *58*(2), 419-70. doi: 10.1257/jel.20191574

Lee, F. S., Pham, X., & Gu, G. (2013). The uk research assessment exercise and the narrowing of uk economics. *Cambridge Journal of Economics*, *37*(4), 693-717. doi: 10.1093/cje/bet031

Mittermaier, B. (2020). Peer review and bibliometrics. In R. Ball (Ed.), *Handbook bibliometrics* (p. 77-90). Berlin/Boston: De Gruyter Saur. doi: 10.1515/9783110646610-009

Rousseau, S., & Rousseau, R. (2021). Bibliometric tecnhniques and their use in business and economics research. *Journal of Economic Surveys*. doi: 10.1111/joes.12415

Stockhammer, E., Dammerer, Q., & Kapur, S. (2021). The research excellence framework 2014, journal ratings and the marginalisation of heterodox economics. *Cambridge Journal of Economics*, *45*(2), 243-269. doi: 10.1093/cje/beaa054

Thomas, D. A., Nedeva, M., Tirado, M. M., & Jacob, M. (2020). Changing research on research evaluation: A critical literature review to revisit the agenda. *Research Evaluation*, *29*(3), 275-288. doi: 10.1093/reseval/rvaa008