On the agreement between bibliometrics and peer review: evidence from the Italian research assessment exercises*

Alberto Baccini† Lucio Barabesi‡ Giuseppe De Nicolao§

Abstract

Two experiments for evaluating the agreement between bibliometrics and informed peer review – based on on two large samples of journal articles – were performed by ANVUR, the Italian governmental agency for research evaluation. They were presented as successful and warranting the combined use of bibliometrics and peer review in research assessment exercises. This paper aims to analyze in full the two experiments and to draw the definitive evidence from them. First, we have provided the correct design-based environment for the inference, since data were collected by means of stratified random sampling. Thus, the design-based estimation of the weighted Cohen's kappa coefficients and the corresponding confidence intervals is developed and adopted for assessing the agreement. In both the experiments, the upper bounds of the confidence intervals for the weighted Cohen's kappa coefficients are smaller – in most cases strictly smaller – than 0.40 (a threshold indicating at most a weak agreement) for each scientific area and for the aggregate data. Therefore, given such a low level of agreement, it is likely that the combined use of bibliometrics and peer review might have introduced uncontrollable major biases in the final results of the Italian research assessment exercises. In addition, as to the second experiment, we have also addressed the problem of missing proportion homogeneity between the scientific areas. We have assessed that the data are missed with unequal proportions between the areas – a further drawback which may invalidate the conclusion carried out by ANVUR. Hence, from the point of view of the academic discussion about the agreement between bibliometrics and peer review, this paper documents that the two ANVUR's experiments do not bring a valid contribution to the debate, since they were designed in a largely unsatisfactory manner.

Keywords: bibliometrics; peer review; research assessment exercise; Cohen's kappa; Gwet's kappa; design-based estimation; stratified random sampling; testing homogeneity of missing proportions.

*Funding: This work was supported by Institute For New Economic Thinking Grant ID INO17-00015.
†Department of Economics and Statistics, University of Siena, Italy; alberto.baccini@unisi.it
‡Department of Economics and Statistics, University of Siena, Italy
§Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy

1 Introduction

The question of the agreement degree between bibliometrics and peer review is the target of an ongoing and growing discussion. The issue is of interest from many different point of view. For example, Research Assessment Designers (RADs) would welcome a definite proof of the agreement of peer review with some kind of bibliometric indicators – since simpler, cheaper and more "objective" bibliometric indicators could wholly replace the peer review (Pride and Peter, 2018). Many RADs and scholars would welcome that definite proof in order to justify the substitution of the unreliable human peer reviewers – prone to nepotism and opportunism – with algorithms. In this perspective, the target is to reach an evaluation without human evaluators, ideally tending to a "view from nowhere" (Goukrager, 2012). It is usual to read about the contraposition of "objective bibliometric data" and "subjective peer reviews". It is not surprising at all that many descriptions of the aims of national research assessments emphasize that point. For example, the second Polish research assessment exercise is "based on a parametric assessment to make the evaluation more objective and independent from its peer" (Kulczycki et al., 2017). Again, in a final report of the research assessment in Italy, some members of the panel claim "the need of strongly reducing the peer evaluation, since it introduces a subjectivity representing a bias that cannot be normalized" (ANVUR, 2017, Area 7 report, p.113). Others RADs would welcome the definite proof of the agreement for a less radical and more practical reason: if bibliometrics and peer review agree, it is possible to combine them in a universal research assessment where some disciplines – notably the humanities – cannot simply be evaluated by indicators.

The biggest researches about the agreement between bibliometrics and peer review were realized in United Kingdom and in Italy at the margin of the national research assessment exercises. The results for the UK and Italy do not converge. In the UK results of the "metric tide" on this respect were synthesized as negative: "This work has shown that individual metrics give significantly different outcomes from the REF peer review process, showing that metrics cannot provide a like-for-like replacement for REF peer review" (Wilsdon et al., 2015).

In Italy, the governmental agency for research evaluation (ANVUR) realized two extensive experiments, which were based on large samples of journal articles. They were conducted during the two national research assessments for the years 2004-2010 (VQR1) and 2011-2014 (VQR2). The two experiments are hereinafter indicated as EXP1 and EXP2, respectively. Results of EXP1 were published not only as official reports, but also disseminated as working papers or scientific papers in refereed journals by scholars working for ANVUR. They were presented as supporting "the choice of using both techniques in order to assess the quality of Italian research institutions" (Ancaiani et al., 2015). Bertocchi and coauthors published five identical working papers where they interpreted the results of EXP1 as claiming that peer review and bibliometrics "are close substitutes" (for all Bertocchi et al., 2013). In the peer reviewed version finally published, they concluded that "the agencies that run these evaluations could feel confident about using bibliometric evaluations and interpret the results as highly correlated with what they would obtain if they performed informed peer review" (Bertocchi et al., 2015). With less hype, the results of EXP2 were also presented as a success in the official report: since a "not-zero correlation" was found (ANVUR, 2017), "we can hence conclude that the combined used of bibliometric indicators for citations and journal impact may provide a useful proxy to assess articles quality" (Alfò et al., 2017).

Two of the authors of the present paper analyzed EXP1. Despite they were unable to access raw data – at the time undisclosed by ANVUR – they documented many flaws in EXP1 (Baccini and De Nicolao, 2016a, 2016b, 2017a, 2017b, Benedetto et al., 2017). After a new request, ANVUR accepted to disclose the anonymous data of both EXP1 and EXP2.[1] Therefore, it is possible to replicate – at least partially[2] – the results of EXP1 and EXP2, by verifying in details ANVUR methods and calculations. As a result, it is now possible to reappraise the evidences about the agreement between bibliometrics and peer review in the Italian research assessment exercises.

The paper is organized as follows: in Section 2, we present the structure of EXP1 and EXP2 by reminding the essential issues of the Italian research assessment exercises. In Section 3, with the aim of re-analyzing the ANVUR data on agreement, we develop a correct framework for the design-based estimation of the Cohen's kappa coefficient. Section 4 presents the estimates of Cohen's kappa coefficients for EXP1 and EXP2, by discussing the current results *vis-à-vis* with the findings given by ANVUR. In Section 5, a further problem with missing data in EXP2 is presented and the homogeneity of missing proportion between scientific areas is assessed. Section 6 concludes by discussing the definitive evidence that can be drawn from the Italian research assessment exercises about the agreement between bibliometrics and peer review.

2 A brief description of the Italian experiments

A brief contextualization of EXP1 and EXP2 is preliminarily needed for understanding their role and relevance in the two Italian research assessments (the present description is largely based on Baccini and De Nicolao, 2016a). The aim of both VQR1 and VQR2 were to evaluate research institutions – such as universities or departments – and research areas and fields, both at national or institutional level. Each university, research institution, department and research field was classified by calculating the average score obtained by the research outputs submitted by researchers. To this end, all the researchers with a permanent position had to submit a fixed number – with few exceptions – of research outputs (3 in VQR1 and 2 in VQR2). Each research work was then evaluated as Excellent (score 1), Good (score 0.8), Acceptable (score 0.5), Limited (score 0) in VQR1, and as Excellent (score 1), Elevated (score 0.7), Fair (score 0.4), Acceptable (score 0.1), Limited (score 0) in VQR2.

Both VQR1 and VQR2 were organized in 16 widely defined research areas. The 16 areas were: Mathematics and Informatics (Area 1), Physics (Area 2), Chemistry (Area 3), Earth Sciences (Area 4), Biology (Area 5), Medicine (Area 6), Agricultural and Veterinary Sciences (Area 7), Civil Engineering and Architecture (Areas 8a and 8b), Industrial and Information Engineering (Area 9), Antiquities, Philology, Literary studies, Art History (Area 10), History, Philosophy, Pedagogy and Psychology (Areas 11a and 11b), Law (Area 12), Economics and Statistics (Area 13), Political and Social Sciences (Area 14). These areas originate from the traditional classification of research areas adopted in Italy. For each area, an evaluation panel was established with a number of panelists proportional to the number of research outputs to be evaluated.

[1] The mail from Alberto Baccini to Prof. Paolo Miccoli (President of ANVUR) containing the request is dated March 12th 2019. The decision of disclosing the data was communicated by mail dated March 26th 2019; the access to the data was open on April 9th 2019.

[2] Replication is possible only at the research Area levels, since – according to a communication dated 16th March 2019 – the data for the sub-Areas "are no longer available" in the ANVUR archives.

Panels directly managed and evaluated subsets of research products submitted for evaluation in their area of expertise. In both research assessments, research evaluation was analogously realized. Panels for the so-called "bibliometric areas" (areas 1-9 excluding area 8a, i.e. hard sciences, engineering and life sciences) evaluated papers mainly – but not exclusively – through bibliometrics. The bibliometric algorithm changed between VQR1 and VQR2, even if in both assessments it was based on the number of citations received by an article and on a journal indicator – e.g. the impact factor – of the journal in which it was published. In the case that the two indicators gave coherent indications, the algorithm generated a score and an evaluation for the article. Otherwise, if they disagreed (i.e. high number of citations and low impact factor or viceversa), the algorithm output was unable to attribute a defined score to the article that was therefore classified as "IR" and evaluated by Informed Peer Review (IPR).

Panels of the so called "non-bibliometric areas" (Areas 8a, 10, 11, 12, 14, i.e. social science and humanities, excluding economics and statistics) evaluated the submitted research products exclusively by IPR. Area 13 (Economics and Statistics) was an exception, since the Area 13 panel developed a journal ranking where journals were classified as Excellent, Good, Acceptable, Limited (in the case of VQR1) or Excellent, Elevated, Fair, Acceptable, Limited (in the case of VQR2). All the articles published in one of the listed journals then received the score of the journal in which they were published. All other research outputs (i.e. books, chapters, articles published in journals not ranked by ANVUR) were evaluated by IPR.

IPR was identically organized in the two research assessments. A publication was assigned to two members of the area panel, who independently chosen two anonymous reviewers, conventionally called P1 and P2. The two reviewers performed the IPR of the article, by using a predefined format (slightly different between the two research assessment and also between panels in the same assessment). Then, the referee reports were received by the two members of the area panel, who formed a so-called Consensus Group (CG) for deciding the final score of the article.

ANVUR coined the expression "evaluative mix" to denote this complex evaluative machinery that created many problems (see e.g. Abramo and D'Angelo, 2016, 2017, Franceschini and Maisano, 2017). The main drawback is the possible bias induced by the adoption of different evaluation techniques. Indeed, if IPR produced scores systematically different from the ones produced by bibliometrics, this might have introduced a systematic bias in the scoring system used for ranking institutions. Indeed, ANVUR precisely realized EXP1 and EXP2 for addressing that problem: a good agreement between the bibliometric evaluation and the evaluation performed by IPR might justify the adoption of the two different evaluation methods and preserve the comparability of results among areas, institutions, departments and research fields. Positive results of EXP1 and EXP2 were crucial for the soundness of Italian research assessment results: if bibliometrics and peer review do not agree and give significantly different results, the average scores of an institution might be distorted by the different percentage of scores attributed by bibliometrics and by IPR.

EXP1 and EXP2 have an identical structure. Both the experiments considered the "bibliometric areas" plus the Area 13, where the "evaluative mix" was applied. The bulk of ANVUR experiments consisted in the analysis of the agreement between the evaluation obtained through IPR and bibliometric algorithms. The rationale of both experiments was very simple: a sample of the journal articles submitted to the research assessment was

scored by IPR and by the bibliometric algorithm. The agreement between the scores is then analyzed by using the Cohen's kappa coefficient (Cohen, 1960), a popular index of interrater agreement for nominal categories (see e.g. Sheskin, 2003). A high level of agreement between IPR and bibliometric scores should be interpreted as justifying the use of both techniques in a same research assessment exercise.

The major difference in the concrete realization of the two experiments was the following. EXP1 was conducted at the same time of the general assessment exercise. This fact had two major consequences. The first was that the reviewers were not able to distinguish between papers evaluated for the experiment and the ones that they had to evaluate for the research assessment. This was true for all research areas with the unique exception of Area 13, where the panelists and the referees knew that all the papers published in journals classified by the panel have to be evaluated for EXP1 (Baccini and De Nicolao, 2016a). The second consequence was that all the papers of the sample were peer reviewed, since the successful completion of the research assessment required the evaluation of all submitted articles. On the contrary, EXP2 started after the conclusion of the research assessment. As a consequence panelists and reviewers were all perfectly aware that they were working for EXP2. Moreover, some papers did not receive a peer review evaluation. Therefore, in EXP2 there were missing data in the sample which were not accounted for by ANVUR when the agreement indexes were computed.

As to the samples for EXP1 and EXP2, ANVUR (2013, 2017, Appendix B) adopted a stratified random sampling with proportional allocation of the population constituted by the journal articles submitted to the research assessments (sample size was about 10% of the population size)[3]. Unfortunately, the final results of both the experiments did not refer to the whole sample of articles but solely to a sub-sample. As we have previously remarked, the bibliometric algorithms might result in an inconclusive classification IR for some articles – for which the disagreement between citations and impact factor did not permit to automatically assign a score. Both in EXP1 and EXP2, all the articles classified as IR were dropped from the experiments, i.e. they received an IPR evaluation, even if they were not considered for the calculation of the agreement. Table 1 and Table 2 shows the sizes of the article population, of the sample and of the reduced sub-sample – according to the stratification based on the areas – for EXP1 and EXP2, respectively. For EXP2 the number of missing papers per area, i.e. papers for which a peer review score was unavailable, is also reported.

For EXP1, the reduction of the sample due to the exclusion of the paper classified as IR was not disclosed neither in ANVUR's official reports nor in Ancaiani et al. (2015). For EXP2, ANVUR proceeded by adopting the same strategy (ANVUR, 2017). Two of the authors of this paper (Baccini and De Nicolao, 2017a), with reference to Ancaiani et al. (2015), raised serious concerns about the whole experiment, by highlighting that unknown biases might have been introduced due to the dropping of the IR items from the sample. In order to gain a basic qualitative intuition of the problems induced by the such a selection of papers, it suffices to observe that ANVUR removed from both EXP1 and EXP2 the more problematic articles for which the bibliometric algorithm was unable to reach a score. We cannot exclude that these articles were also the more problematic to be evaluated by peer

[3]Indeed, the Final Reports remark that: "The sample was stratified according to the distribution of the products among the sub-areas of the various areas" (ANVUR, 2017, Appendix B, p.1 our translation). For EXP1 results were published at a sub-area level, while for EXP2 results were solely published for areas. Moreover, the raw data at the sub-area level are not yet available.

Table 1: Population, sample and sub-sample sizes for scientific areas in EXP1.

Scientific Areas	Population	Sample	Sub-sample
Area 1 - Mathematics and Informatics	6758	631	438
Area 2 - Physics	15029	1412	1212
Area 3 - Chemistry	10127	927	778
Area 4 - Earth Sciences	5083	458	377
Area 5 - Biology	14043	1310	1058
Area 6 - Medicine	21191	1984	1602
Area 7 - Agricultural and Veterinary Sciences	6284	532	425
Area 8a - Civil Engineering	2460	225	198
Area 9 - Industrial and Information Engineering	12349	1130	919
Area 13 - Economics and Statistics	5681	590	590
	99005	9199	7597

Source: ANVUR (2013, Appendix B).

Table 2: Population, sample, sub-sample sizes and number of missing articles for scientific areas in EXP2.

Scientific Areas	Population	Sample	Sub-sample	Missing
Area 1 - Mathematics and Informatics	4631	467	344	23
Area 2 - Physics	10182	1018	926	10
Area 3 - Chemistry	6625	662	549	9
Area 4 - Earth Sciences	3953	394	320	6
Area 5 - Biology	10423	1037	792	86
Area 6 - Medicine	15400	1524	1071	231
Area 7 - Agricultural and Veterinary Sciences	6354	638	489	8
Area 8b - Civil Engineering	2370	237	180	3
Area 9 - Industrial and Information Engineering	9930	890	739	108
Area 11b - Psychology	1801	180	133	5
Area 13 - Economics and Statistics	5490	512	498	14
	77159	7667	6041	503

Source: ANVUR (2017, Appendix B).

reviewers. If this issue were true, ANVUR calculated the agreement indicator on sub-samples "more favorable" to agreement than the complete samples. Moreover, in EXP2 there also occurs a problem with missing articles: further drawbacks would arise if the distribution of missing articles in the sample occurred in a non-proportional way between the strata.

On the basis of the previous discussion, in the following section the design-based estimation of the weighted Cohen's kappa coefficient is developed in view of defining the correct inferential setting for the problem of the agreement between bibliometrics and peer review in EXP1 and EXP2.

3 Design-based estimation of the Cohen's kappa coefficient

Let us assume a fixed population of N items which are classified into c categories on the basis of two ratings. For the sake of simplicity, the c categories are labeled on the set $I = \{1, \ldots, c\}$. Hence, the j-th item of the population is categorized according to the first evaluation – say $u_j \in I$ – and the second evaluation – say $v_j \in I$ – for $j = 1, \ldots, N$.

A commonly-adopted measure of agreement between classifications of two raters is given by the Cohen's kappa coefficient and its weighted generalization (Cohen, 1960, 1968). Practitioners often adopt this index in order to assess the inter-rater agreement for categorical ratings, while its weighted counterpart is preferred when the categories can be considered ordinal (see e.g. Fagerland et al., 2017, p.548, Berry et al., 2018, p.596). However, the Cohen's kappa coefficient is also criticized for some methodological drawbacks (for more details, see Strijbos et al., 2006, and Uebersax, 1987, among others).

In the present design-based approach, the "population" weighted Cohen's kappa coefficient is defined as

$$\kappa_w = \frac{p_o - p_e}{1 - p_e} \ , \tag{1}$$

where

$$p_o = \sum_{l=1}^{c} \sum_{m=1}^{c} w_{lm} p_{lm} \ , \ p_e = \sum_{l=1}^{c} \sum_{m=1}^{c} w_{lm} p_{l+} p_{+m} \ ,$$

while

$$p_{lm} = \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}_{\{l\}}(u_j) \mathbf{1}_{\{m\}}(v_j) \ , \ p_{l+} = \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}_{\{l\}}(u_j) \ , \ p_{+l} = \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}_{\{l\}}(v_j)$$

and $\mathbf{1}_B$ is the usual indicator function of a set B, i.e. $\mathbf{1}_B(u) = 1$ if $u \in B$ and $\mathbf{1}_B(u) = 0$, otherwise. In practice, p_{lm} is the proportion of items in the population classified into the l-th category according to the first rating and into the m-th category according to the second rating. Similarly, p_{l+} and p_{+l} are the proportions of items categorized into the l-th category according to the first rating and the second rating, respectively. In addition, the weights w_{lm}, with $l, m = 1, \ldots, c$, are suitably chosen in order to consider the magnitude of disagreement (see e.g. Fagerland et al., 2017, p.551). In particular, the (usual) unweighted Cohen's kappa coefficient is obtained from (1) when $w_{lm} = 1$ if $l = m$ and $w_{lm} = 0$, otherwise. It is worth remarking that, for estimation purposes, we have conveniently expressed the Cohen's kappa coefficient (1) as a smooth function of population totals – i.e. the p_{lm}'s, the p_{l+}'s and the p_{+l}'s.

Let us now assume that a sampling design is adopted in order to estimate κ_w and let us consider a sample of fixed size n. Moreover, let S denote the set of indexes corresponding to the sampled items – i.e. a subset of size n of the first N integers – and let π_j be the inclusion probability of the first order for the j-th item. As an example aimed to the subsequent application, let us assume that the population be partitioned into L strata and that N_h is the size of the h-th stratum with $h = 1, \ldots, L$. Obviously, it holds $N = \sum_{h=1}^{L} N_h$. If a stratified sampling design is considered, the sample is obtained by drawing n_h items in the l-th stratum by means of simple random sampling without replacement in such a way that $n = \sum_{h=1}^{L} n_h$. Therefore, as is well known, it turns out that $\pi_j = n_h/N_h$ if the j-th item is in the h-th stratum (see e.g. Thompson, 1997). When a proportional allocation is adopted, it also holds that $n_h = nN_h/N$ – and hence it obviously follows $\pi_j = n/N$.

In order to obtain the estimation of κ_w, it is should noticed that

$$\widehat{p}_{lm} = \frac{1}{N} \sum_{j \in S} \frac{\mathbf{1}_{\{l\}}(u_j) \mathbf{1}_{\{m\}}(v_j)}{\pi_j} \ , \ \widehat{p}_{l+} = \frac{1}{N} \sum_{j \in S} \frac{\mathbf{1}_{\{l\}}(u_j)}{\pi_j} \ , \ \widehat{p}_{+l} = \frac{1}{N} \sum_{j \in S} \frac{\mathbf{1}_{\{l\}}(v_j)}{\pi_j} \ ,$$

are unbiased Horvitz-Thompson estimators of the population proportions p_{lm}, p_{l+} and p_{+l}, respectively. Thus, by bearing in mind the general comments provided by Demnati and Rao (2004) on the estimation of a function of population totals, a "plug-in" estimator of (1) is given by

$$\widehat{\kappa}_w = \frac{\widehat{p}_o - \widehat{p}_e}{1 - \widehat{p}_e} \ , \tag{2}$$

where

$$\widehat{p}_o = \sum_{l=1}^{c} \sum_{m=1}^{c} w_{lm} \widehat{p}_{lm} \ , \ \widehat{p}_e = \sum_{l=1}^{c} \sum_{m=1}^{c} w_{lm} \widehat{p}_{l+} \widehat{p}_{+m} \ .$$

Even if estimator (2) is biased, its bias is negligible since (1) is a differentiable function of the population totals with non-null derivatives (for more details on such a result, see e.g. Thompson, 1997, p.106).

As usual, variance estimation is mandatory is order to achieve an evaluation of the accuracy of the estimator. Since (2) is a rather involved function of sample totals, its variance may be conveniently estimated by the linearization method or by the jackknife technique (see e.g. Demnati and Rao, 2004, and references therein). Alternatively, a bootstrap approach – which is based on a pseudo-population method – may be suitably considered (for more details on this topic, see e.g. Quatember, 2015).

It should be remarked that inconclusive ratings occur in EXP1 and EXP2 and – in addition – missing ratings are also present in EXP2. However, even if ANVUR does not explicitly states this issue, its target seems to be the sub-population of items with two reported ratings. Hence, some suitable variants of the Cohen's kappa coefficient have to be considered. In order to deal with an appropriate definition of the population parameter in this setting, the three suggestions provided by De Raadt et al. (2019) could be adopted. For the sake of simplicity, let us suppose that inconclusive or missing ratings are classified into the c-th category. A first way to manage the issue consists in deleting all items which are not classified by both raters and apply the weighted Cohen's kappa coeffcient to the items with two ratings (see also Strijbos and Stahl, 2007). After some straightforward algebra, this variant of the population weighted Cohen's kappa coefficient may be written as

$$\kappa_w^{(1)} = \frac{p_o^{(1)} - p_e^{(1)}}{1 - p_e^{(1)}} \ , \tag{3}$$

where

$$p_o^{(1)} = \frac{\sum_{l=1}^{c-1}\sum_{m=1}^{c-1} w_{lm} p_{lm}}{\sum_{l=1}^{c-1}\sum_{m=1}^{c-1} p_{lm}} \ , \ p_e^{(1)} = \frac{\sum_{l=1}^{c-1}\sum_{m=1}^{c-1} w_{lm}(p_{l+} - p_{lc})(p_{+m} - p_{cm})}{(\sum_{l=1}^{c-1}\sum_{m=1}^{c-1} p_{lm})^2} \ .$$

It is worth noting that (3) could be not a satisfactory index, since it does not take into account the size of inconclusive or missing ratings. Similarly to (1), its variant (3) can be estimated as

$$\widehat{\kappa}_w^{(1)} = \frac{\widehat{p}_o^{(1)} - \widehat{p}_e^{(1)}}{1 - \widehat{p}_e^{(1)}} \ , \tag{4}$$

where

$$\widehat{p}_o^{(1)} = \frac{\sum_{l=1}^{c-1}\sum_{m=1}^{c-1} w_{lm} \widehat{p}_{lm}}{\sum_{l=1}^{c-1}\sum_{m=1}^{c-1} \widehat{p}_{lm}} \ , \ \widehat{p}_e^{(1)} = \frac{\sum_{l=1}^{c-1}\sum_{m=1}^{c-1} w_{lm}(\widehat{p}_{l+} - \widehat{p}_{lc})(\widehat{p}_{+m} - \widehat{p}_{cm})}{(\sum_{l=1}^{c-1}\sum_{m=1}^{c-1} \widehat{p}_{lm})^2} \ .$$

The second proposal by De Raadt et al. (2019) for a variant of the weighted Cohen's kappa coefficient is based on Gwet's kappa (Gwet, 2014). The population weighted Gwet's kappa may be defined as

$$\kappa_w^{(2)} = \frac{p_o^{(2)} - p_e^{(2)}}{1 - p_e^{(2)}} \ , \tag{5}$$

where

$$p_o^{(2)} = \frac{\sum_{l=1}^{c-1}\sum_{m=1}^{c-1} w_{lm} p_{lm}}{\sum_{l=1}^{c-1}\sum_{m=1}^{c-1} p_{lm}} \ , \ p_e^{(2)} = \frac{\sum_{l=1}^{c-1}\sum_{m=1}^{c-1} w_{lm} p_{l+} p_{+m}}{(1 - p_{c+})(1 - p_{+c})} \ .$$

This index considers the sizes of inconclusive or missing ratings. Indeed, even if $p_o^{(2)} = p_o^{(1)}$, the quantity $p_e^{(2)}$ is actually a weighted sum of the products of type $p_{l+}p_{+l}$ – in contrast to the quantity $p_e^{(1)}$ which is a weighted sum of the products of type $(p_{l+} - p_{lc})(p_{+m} - p_{cm})$. In turn, (5) may be estimated by means of

$$\widehat{\kappa}_w^{(2)} = \frac{\widehat{p}_o^{(2)} - \widehat{p}_e^{(2)}}{1 - \widehat{p}_e^{(2)}} \ , \tag{6}$$

where

$$\widehat{p}_o^{(2)} = \frac{\sum_{l=1}^{c-1}\sum_{m=1}^{c-1} w_{lm}\widehat{p}_{lm}}{\sum_{l=1}^{c-1}\sum_{m=1}^{c-1}\widehat{p}_{lm}} \ , \ \widehat{p}_e^{(2)} = \frac{\sum_{l=1}^{c-1}\sum_{m=1}^{c-1} w_{lm}\widehat{p}_{l+}\widehat{p}_{+m}}{(1-\widehat{p}_{c+})(1-\widehat{p}_{+c})} \ .$$

The third proposal by De Raadt et al. (2019) for a variant of (1) stems on assuming null weights for the inconclusive or missing ratings, i.e. by assuming that $w_{lm} = 0$ if $l = c$ or $m = c$. Hence, this variant is obviously defined as

$$\kappa_w^{(3)} = \frac{p_o^{(3)} - p_e^{(3)}}{1 - p_e^{(3)}} \ , \tag{7}$$

where

$$p_o^{(3)} = \sum_{l=1}^{c-1}\sum_{m=1}^{c-1} w_{lm}p_{lm} \ , \ p_e^{(3)} = \sum_{l=1}^{c-1}\sum_{m=1}^{c-1} w_{lm}p_{l+}p_{+m} \ .$$

In turn, (7) may be estimated by means of

$$\widehat{\kappa}_w^{(3)} = \frac{\widehat{p}_o^{(3)} - \widehat{p}_e^{(3)}}{1 - \widehat{p}_e^{(3)}} \ , \tag{8}$$

where

$$\widehat{p}_o^{(3)} = \sum_{l=1}^{c-1}\sum_{m=1}^{c-1} w_{lm}\widehat{p}_{lm} \ , \ \widehat{p}_e^{(3)} = \sum_{l=1}^{c-1}\sum_{m=1}^{c-1} w_{lm}\widehat{p}_{l+}\widehat{p}_{+m} \ .$$

The previous findings are applied to the data collected in EXP1 and EXP2 in the following section.

4 Cohen's kappa coefficient estimation in the Italian experiment

On the basis of the comments given in Section 3, we have considered the estimation of the Cohen's kappa coefficient for the agreement between the bibliometric and the peer ratings, as well as for the agreement between the ratings of the first referee (P1) and the second referee (P2). ANVUR has adopted a stratified sampling design for the two populations of articles in EXP1 and EXP2 of sizes $N = 99{,}005$ and $N = 77{,}159$, respectively. The sizes of the strata in EXP1 and EXP2 – as well as the corresponding sample sizes – are reported in Tables 1 and 2.

The choice of the weights for the calculation of Cohen's kappa coefficient is subjective – indeed, the adoption of different sets of weights may modify the results on the agreement. In order to reproduce ANVUR's results, we solely use the sets of the so-called VQR-weights adopted by ANVUR. By assuming that $\mathbf{W} = (w_{l,m})$ generally denote the matrix of weights,

the matrices of the VQR-weights for EXP1 and EXP2 are respectively given by

$$\mathbf{W} = \begin{pmatrix} 1 & 0.8 & 0.5 & 0 \\ 0.8 & 1 & 0.7 & 0.2 \\ 0.5 & 0.7 & 1 & 0.5 \\ 0 & 0.2 & 0.5 & 1 \end{pmatrix}$$

and

$$\mathbf{W} = \begin{pmatrix} 1 & 0.7 & 0.4 & 0.1 & 0 \\ 0.7 & 1 & 0.7 & 0.4 & 0.3 \\ 0.4 & 0.7 & 1 & 0.7 & 0.6 \\ 0.1 & 0.4 & 0.7 & 1 & 0.9 \\ 0 & 0.3 & 0.6 & 0.9 & 1 \end{pmatrix}.$$

The two system of weights were based on the score adopted in the research assessment, even if they appear as counter-intuitive since they attribute different weights to a same category distance. For example, in EXP1 a distance of two categories is weighted 0.5 if it is realized between the first and the third category, while it is solely weighted 0.2 if it is realized between second and fourth category.

At first, we have considered the estimation (3),(5) and (7) for the agreement of the bibliometric and peer review ratings by means of the estimators (4), (6) and (8). The estimation was carried out for each area and for the global population in both EXP1 and EXP2. Variance estimation was carried out by means of the Horvitz-Thompson based bootstrap – stemming on the use of a pseudo-population – which is described by Quatember (2015, p.16, p.80). The whole computation was implemented by means of the algebraic software Mathematica (Wolfram Research Inc., 2014).

The point estimates and the corresponding confidence intervals are given in Tables 3 and 4, reporting also the point estimates and the confidence intervals provided by ANVUR (2013, 2017). It seems that ANVUR considered the estimation of (3), even if this issue is not explicitly stated in its reports (ANVUR, 2013, 2017). Actually, the point estimates given by ANVUR correspond to those computed by means of (4), so that the estimated Cohen's kappa coefficient did not account for the size of inconclusive ratings in EXP1 and for the size of inconclusive or missing ratings in EXP2. However, even if ANVUR has apparently adopted a design-based inference, the variance estimation is carried out in a model-based approach. As a matter of fact, the confidence intervals provided by ANVUR are the same computed by means of the packages **psych** and **vcd** of the software R (R Core Team, 2019) in the case of EXP1 and EXP2, respectively. Unfortunately, these confidence intervals rely on the model-based approximation for large samples described by Fleiss et al. (2003, p.610).

In the case of (4), the bootstrap method generally produces confidence intervals which are narrower than those computed by ANVUR - consistently with the fact that a stratified sampling design is carried out, rather than a simple random sampling design. In the cases of (6) and (8), the point estimates are respectively larger and smaller than those obtained by (4). The upper bounds of the confidence intervals are generally smaller than 0.40 in Tables 3 and 4 for (6), and smaller than 0.30 for (8). The unique exception is Area 13 in EXP1, where the experiment was substantially modified with respect to the other areas (Baccini and De Nicolao, 2016a, 2016b). A unweighted Cohen's kappa coefficient in the range $(0.20, 0.40)$ can be considered a rather weak value of agreement – see e.g. the recent

Table 3: Cohen's kappa coefficient estimates (percent) for EXP1 (95% confidence level intervals in parenthesis), bibliometric vs peer review ratings.

Area	ANVUR[1]	$\widehat{\kappa}_w^{(1)}$	$\widehat{\kappa}_w^{(2)}$	$\widehat{\kappa}_w^{(3)}$
1	31.73(23.00,40.00)	31.73(25.21,38.26)	33.40(26.80,40.00)	15.07(11.76,18.38)
2	25.15(21.00,29.00)	25.15(21.10,29.19)	29.15(25.29,33.01)	18.91(16.24,21.58)
3	22.96(17.00,29.00)	22.96(18.05,27.86)	23.98(19.09,28.88)	14.52(11.32,17.71)
4	29.85(21.00,39.00)	29.85(23.32,36.37)	30.24(23.69,36.79)	20.32(15.66,24.99)
5	34.53(29.00,40.00)	34.53(30.51,38.54)	36.62(32.72,40.51)	23.85(21.13,26.58)
6	33.51(29.00,38.00)	33.51(30.30,36.72)	34.62(31.47,37.77)	22.73(20.51,24.95)
7	34.37(27.00,42.00)	34.37(27.99,40.75)	36.62(30.59,42.65)	22.60(18.43,26.77)
8a	22.61(11.00,34.00)	22.61(12.70,32.52)	22.99(13.06,32.92)	16.35(8.90,23.80)
9	17.10(13.00,21.00)	17.10(13.17,21.03)	21.95(17.78,26.11)	12.56(10.12,15.01)
13	61.04(53.00,69.00)[2]	54.17(49.37,58.98)	54.17(49.37,58.98)	54.17(49.37,58.98)
All	38.00(36.00,40.00)[3]	34.15(32.64,35.66)	35.76(34.28,37.24)	23.28(22.23,24.33)

[1] Source: ANVUR (2013, Appendix B). Reproduced in Ancaiani et al. (2015).
[2] Estimate with the wrong system of weights as documented in Baccini and De Nicolao (2017a). Benedetto et al. (2017) justified it as "factual error in editing of the table" and published a corrected estimate of 54.17.
[3] Ancaiani et al. (2015) reported a different estimate of 34.41, confirmed also in Benedetto et al. (2017).

Table 4: Cohen's kappa coefficient estimates (percent) for EXP2 (95% confidence level intervals in parenthesis), bibliometric vs peer review ratings.

Area	ANVUR[1]	$\widehat{\kappa}_w^{(1)}$	$\widehat{\kappa}_w^{(2)}$	$\widehat{\kappa}_w^{(3)}$
1	21.50(15.10,27.80)	21.48(15.38,27.58)	22.85(16.71,29.00)	14.97(11.79,18.16)
2	26.50(22.40,30.50)	26.48(22.61,30.34)	28.66(24.86,32.46)	22.35(19.46,25.23)
3	19.50(14.30,24.70)	19.49(14.60,24.38)	20.85(16.01,25.69)	13.71(10.71,16.72)
4	23.90(16.60,31.20)	23.90(17.02,30.77)	24.52(17.75,31.28)	15.78(11.55,20.01)
5	24.10(19.70,28.40)	24.07(19.98,28.15)	25.01(20.97,29.05)	19.93(17.75,22.11)
6	22.80(19.50,26.20)	22.83(19.62,26.04)	24.47(21.32,27.62)	21.00(19.49,22.51)
7	27.00(21.30,32.70)	27.01(21.66,32.36)	28.76(23.56,33.96)	16.02(13.05,18.99)
8b	17.20(8.80,25.60)	17.21(9.183,25.23)	20.36(12.55,28.16)	11.45(7.23,15.67)
9	16.90(12.90,21.00)	16.91(13.04,20.78)	19.62(15.82,23.42)	18.51(16.58,20.44)
11b	24.10(13.70,34.50)	24.09(14.30,33.88)	25.45(15.93,34.97)	14.76(9.556,19.95)
13	30.90(26.20,35.50)	30.85(26.36,35.34)	30.85(26.36,35.34)	31.54(27.51,35.57)
All	26.00(24.50,27.60)	26.10(24.64,27.56)	27.31(25.87,28.74)	20.88(20.05,21.71)

[1] Source: ANVUR (2017, Appendix B).

guideline for interpreting Cohen's kappa coefficient by Fagerland et al. (2017, p.549) and the survey provided by Baccini and De Nicolao (2016a). Hence, a weighted Cohen's kappa coefficient in this interval is an even worse value of agreement. In addition, Baccini and De Nicolao (2016a) showed that the system of weights adopted by ANVUR is the most favourable, i.e. it produced the highest values of agreement, with respect to alternative system of weights – e.g. linear – or to unweighted Cohen's kappa coefficient.

Subsequently, we have considered the estimation of the Cohen's kappa coefficient for the agreement of the ratings of P1 and P2 in EXP1 and EXP2, respectively. As to EXP1, we have considered the estimation of (1) in the whole population and the estimation of the same index for the two sub-populations which received a definite bibliometric rating or an inconclusive bibliometric rating, respectively. It should be remarked that there are not inconclusive bibliometric ratings for Area 13. ANVUR seems to aim estimating (1) in the sub-population bearing a definite bibliometric rating, even if not explicitly stated in its report (ANVUR, 2013) and this issue is actually confirmed in Table 5. By means of an analysis of Table 5, it is apparent that the point estimates provided by ANVUR are

Table 5: Cohen's kappa coefficient estimates (percent) for EXP1 (95% confidence level intervals in parenthesis), P1 vs P2 ratings.

Area	ANVUR[1]	$\widehat{\kappa}_w$	$\widehat{\kappa}_w$ DBR[5]	$\widehat{\kappa}_w$ IPR[6]
1	35.16(26.00,44.00)	33.31(27.54,39.09)	35.16(25.26,45.06)	28.87(18.17,39.57)
2	22.71(18.00,28.00)	23.42(19.44,27.41)	22.71(17.28,28.14)	19.31(9.227,29.39)
3	23.81(17.00,30.00)	20.83(16.00,25.65)	23.81(17.73,29.89)	2.56(-7.01,12.15)
4	25.48(15.00,36.00)	23.27(16.55,30.00)	25.48(16.59,34.36)	12.37(-3.47,28.23)
5	27.17(21.00,33.00)	24.85(20.76,28.93)	27.17(21.56,32.78)	11.12(1.77,20.46)
6	23.56(19.00,29.00)	21.85(18.57,25.12)	23.56(19.09,28.03)	11.84(4.19,19.48)
7	26.56(21.00,33.00)[2]	17.47(11.34,23.61)	16.99(8.15,25.83)	16.41(2.91,29.90)
8a	19.43(6.00,32.00)	19.92(9.64,30.21)	19.43(6.65,32.20)	23.77(-7.45,54.99)
9	18.18(12.00,24.00)	19.39(14.93,23.84)	18.18(11.72,24.64)	21.1(10.70,31.50)
13	45.99(38.00,54.00)[3]	38.98(33.50,44.47)	38.98(33.50,44.47)	-
All	33.00(31.00,35.00)[4]	26.68(25.16,28.20)	27.92(25.90,29.95)	18.90(15.30,22.50)

[1] Source: ANVUR (2013, Appendix B). Reproduced in Ancaiani et al. (2015).
[2] Estimate with the wrong system of weights reported in Baccini and De Nicolao (2017a). Benedetto et al. (2017) justified it as "factual error in editing of the table" and published a corrected estimate of 16.99.
[3] Estimate with the wrong system of weights reported in Baccini and De Nicolao (2017a). Benedetto et al. (2017) justified it as "factual error in editing of the table" and published a corrected estimate of 38.998.
[4] Ancaiani et al. (2015) reported a different estimate of 28.16, confirmed also in Benedetto et al. (2017).
[5] Weighted Cohen's kappa for the sets of articles with a definite bibliometric rating (DBR).
[6] Weighted Cohen's kappa for the sets of articles without a definite bibliometric rating and submitted to informed peer review (IPR).

similar to those based on (2) for the sub-population with a definite bibliometric rating. It is also apparent that they are generally greater than those for the sub-population with an inconclusive bibliometric rating. This last issue confirms the intuition that articles for which bibliometrics rating was inconclusive were also the articles more difficult to evaluate for reviewers who, in fact, had a lesser degree of agreement for these papers. In turn, the confidence intervals provided by ANVUR are the same computed by means of the package **psych** of the software R (R Core Team, 2019) and hence the previous comments also apply in this setting. As to EXP2, missing peer ratings occur and hence the target parameter is given by (3). Similarly to EXP1, we have considered the estimation of (3) also in the case of the two sub-populations and the results are given in Table 6. In turn, the point estimates based on (4) for the population are similar to those for the sub-population with a definite bibliometric rating. However, in this case they are not larger than those for the sub-population with an inconclusive bibliometric rating, even if these point estimates are small – and in some cases very small.

5 Testing homogeneity of missing proportions between strata

In the case of EXP2, we have considered in Section 4 the sizes of missing peer ratings as fixed and – accordingly – we have carried out a design-based approach for the estimation of rating agreement. However, it could be also interesting to assess the homogeneity of missing proportions in the different areas by assuming a random model for the missing peer ratings, i.e. by considering a model-based approach for missing proportion estimation and testing. In order to provide an appropriate setting in such a case, let us suppose in turn a fixed population of N items partitioned into L strata. Moreover, a stratified sampling design is adopted and the notations introduced in Section 2 are assumed. Hence, each item in the h-th stratum may be missed with probability $\theta_h \in [0, 1]$ – independently with respect to

Table 6: Cohen's kappa coefficient estimates (percent) for EXP2 (95% confidence level intervals in parenthesis), P1 vs P2 rating.

Area	ANVUR[1]	$\widehat{\kappa}_w^{(1)}$	$\widehat{\kappa}_w^{(1)}$ DBR[2]	$\widehat{\kappa}_w^{(1)}$ IPR[3]
1	20.20(12.90,27.50)	23.92(17.68,30.15)	20.18(11.11,29.25)	35.71(22.11,49.31)
2	19.50(14.60,24.40)	21.13(16.75,25.52)	19.50(14.24,24.77)	20.29(6.81,33.77)
3	14.00(7.90,20.10)	14.67(9.36,19.98)	13.99(7.14,20.83)	15.53(3.04,28.02)
4	18.90(11.10,26.80)	18.63(11.97,25.29)	18.94(10.14,27.75)	12.4(-2.21,27.02)
5	19.50(14.60,24.50)	20.21(15.80,24.63)	19.53(13.65,25.41)	20.73(9.82,31.63)
6	19.10(17.90,23.20)	17.84(14.24,21.44)	19.08(14.29,23.87)	7.69(-0.73,16.13)
7	19.60(13.40,25.80)	22.38(17.14,27.63)	19.57(11.58,27.57)	28.34(17.54,39.15)
8b	3.50(-0.06,13.20)	8.70(0.22,17.19)	3.47(-9.42,16.37)	22.41(5.90,38.92)
9	15.10(9.90,20.30)	15.36(10.84,19.89)	15.09(8.87,21.31)	12.71(1.65,23.76)
11b	25.70(13.30,38.20)	25.79(15.39,36.19)	25.72(8.93,42.50)	20.68(0.22,41.15)
13	31.20(25.40,36.90)	31.15(25.69,36.61)	31.15(25.69,36.61)	-
All	23.40(21.60,25.20)	23.54(21.97,25.10)	23.50(21.48,25.52)	19.85(15.94,23.77)

[1] Source: ANVUR (2017).
[2] Weighted Cohen's kappa for the sets of articles with a definite bibliometric rating (DBR).
[3] Weighted Cohen's kappa for the sets of articles without a definite bibliometric rating and submitted to informed peer review (IPR).

the other items. Thus, the size of missing items in the h-th stratum, say M_h, is a random variable (r.v.) distributed according to the Binomial law with parameters N_h and θ_h, i.e. the probability function (p.f.) of M_h turns out to be

$$p_{M_h}(m) = \binom{N_h}{m} \theta_h^m (1-\theta_h)^{N_h - m} \mathbf{1}_{\{0,1,\ldots,N_h\}}(m) \ .$$

Let us assume that the r.v. X_h represents the size of missing items of the h-th stratum in the sample. By supposing that the items are missing independently with respect to the sampling design, the distribution of the r.v. X_h given the event $\{M_h = m\}$ is the Hypergeometric law with parameters n_h, m and N_h, i.e. the corresponding conditioned p.f. is given by

$$p_{X_h|\{M_h=m\}}(x) = \frac{\binom{m}{x}\binom{N_h-m}{n_h-x}}{\binom{N_h}{n_h}} \mathbf{1}_{\{\max(0,n_h-N_h+m),\ldots,\min(n_h,m)\}}(x) \ .$$

Hence, on the basis of this finding and by using the result by Johnson et al. (2005, p.377), the r.v. X_h is distributed according to the Binomial law with parameters n_h and θ_h, i.e. the p.f. of X_h is

$$p_{X_h}(x) = \binom{n_h}{x} \theta_h^x (1-\theta_h)^{n_h - x} \mathbf{1}_{\{0,1,\ldots,n_h\}}(x)$$

for each $h = 1, \ldots, L$. Obviously, the X_h's are independent r.v.'s.

Under the frequentist paradigm, let us consider the null hypothesis of missing proportion homogeneity $H_0 : \theta_h = \theta, \forall h = 1, \ldots, L$, versus the alternative hypothesis $H_1 : \theta_h \neq \theta, \exists h = 1, \ldots, L$. For a given $(x_1, \ldots, x_L) \in \mathbb{N}^L$ such that $y = \sum_{h=1}^L x_h$, the likelihood function under the null hypothesis is given by

$$L_0(\theta) \propto \theta^y (1-\theta)^{n-y} \mathbf{1}_{[0,1]}(\theta) \ ,$$

while the likelihood function under the alternative hypothesis is given by

$$L_1(\theta_1, \ldots, \theta_L) \propto \prod_{h=1}^L \theta_h^{x_h} (1-\theta_h)^{n_h - x_h} \mathbf{1}_{[0,1]^L}(\theta_1, \ldots, \theta_L) \ .$$

Thus, the likelihood estimator of θ under the null hypothesis turns out to be $\widehat{\theta} = Y/n$, where $Y = \sum_{h=1}^{L} X_h$. In addition, the likelihood estimator of $(\theta_1, \ldots, \theta_L)$ under the alternative hypothesis turns out to be $(\widehat{\theta}_1, \ldots, \widehat{\theta}_L)$, where $\widehat{\theta}_h = X_h/n_h$.

The likelihood-ratio test statistic could be adopted in order to assess the null hypothesis. However, in the present setting the large-sample results are precluded since the sample size n is necessarily bounded by N and the data sparsity could reduce the effectiveness of the large-sample approximations. A more productive approach may be based on conditional testing (see e.g. Lehmann and Romano, 2005, Chapter 10). First, we consider the χ^2 test statistic – asymptotically equivalent in distribution to the likelihood-ratio test statistic – which in this case, after some algebra, reduces to

$$R := R(X_1, \ldots, X_L) = \sum_{h=1}^{L} \frac{n_h(\widehat{\theta}_h - \widehat{\theta})^2}{\widehat{\theta}(1-\widehat{\theta})} \ .$$

It should be remarked that the r.v. Y is sufficient for θ under the null hypothesis. Hence, the distribution of the random vector $(X_1, \ldots, X_L)$ given the event $\{Y = y\}$ does not depend on θ. Moreover, under the null hypothesis, the distribution of the random vector $(X_1, \ldots, X_L)$ given the event $\{Y = y\}$ is the multivariate Hypergeometric law with parameters y and $(n_1, \ldots, n_L)$, i.e. the corresponding conditioned p.f. turns out to be

$$p_{(X_1,\ldots,X_L)|\{Y=y\}}(x_1, \ldots, x_L) = \frac{\prod_{h=1}^{L} \binom{n_h}{x_h}}{\binom{n}{y}} \mathbf{1}_A(x_1, \ldots, x_L) \ ,$$

where $A = \{(x_1, \ldots, x_L) \in \mathbb{N}^L : x_h \in \{\max(0, n_h - n + y), \ldots, \min(n_h, y)\}, \sum_{h=1}^{L} x_h = y\}$. Thus, by assuming the conditional approach, an exact test may be carried out. Indeed, if r represents the observed realization of the test statistic R, the corresponding P-value turns out to be

$$P(R \geq r \mid \{Y = y\}) = \sum_{(x_1,\ldots,x_L) \in C_r} p_{(X_1,\ldots,X_L)|\{Y=y\}}(x_1, \ldots, x_L) \ ,$$

where $C_r = \{(x_1, \ldots, x_L) \in A : R(x_1, \ldots, x_L) \geq r\}$. It should be remarked that the previous P-value may be approximated by means of a Monte Carlo method by generating realizations of a Hypergeometric random vector with parameters y and $(n_1, \ldots, n_L)$. The generation of each realization requires $(L-1)$ Hypergeometric random variates – for which suitable algorithms exist – and hence the method is practically feasible.

Under the Bayesian paradigm, the missing probability homogeneity between strata may be specified as the model $\mathcal{M}_0$ which assumes that X_l be distributed according to the Binomial law with parameters n_l and θ, for $l = 1, \ldots, L$. In contrast, the model $\mathcal{M}_1$ under the general alternative postulates that X_l be distributed according to the Binomial law with parameters n_l and θ_l, for $l = 1, \ldots, L$. By assuming prior distributions in such a way that θ is elicited as the absolutely-continuous r.v. Θ defined on $[0, 1]$ with probability density function (p.d.f.) given by f_Θ, while $(\theta_1, \ldots, \theta_L)$ is elicited as the vector $(\Theta_1, \ldots, \Theta_L)$ of absolutely-continuous r.v.'s defined on $[0, 1]^L$ with joint p.d.f. given by $f_{\Theta_1,\ldots,\Theta_L}$, the Bayes

factor may be given by

$$B_{1,0} = \frac{\int_{[0,1]^L} \left\{ \prod_{l=1}^{L} \binom{n_l}{x_l} \theta_l^{x_l} (1-\theta_l)^{n_l - x_l} \right\} f_{\Theta_1,\ldots,\Theta_L}(\theta_1,\ldots,\theta_L) d\theta_1 \ldots d\theta_L}{\int_{[0,1]} \left\{ \prod_{l=1}^{L} \binom{n_l}{x_l} \theta^{x_l} (1-\theta)^{n_l - x_l} \right\} f_{\Theta}(\theta) d\theta}$$
$$= \frac{\int_{[0,1]^L} \prod_{l=1}^{L} \theta_l^{x_l} (1-\theta_l)^{n_l - x_l} f_{\Theta_1,\ldots,\Theta_L}(\theta_1,\ldots,\theta_L) d\theta_1 \ldots d\theta_L}{\int_{[0,1]} \theta^{y} (1-\theta)^{n-y} f_{\Theta}(\theta) d\theta} .$$

If conjugate priors are considered, the r.v. Θ is assumed distributed according to the Beta law with parameters a and b, while $(\Theta_1, \ldots, \Theta_L)$ is the vector of r.v.'s with independent components, in such a way that each Θ_l is distributed according to the Beta law with parameters a_l and b_l. It is worth noting that – in a similar setting – a slightly general hierarchical model is considered by Kass and Raftery (1995) (see also Albert, 2009, p.190). Hence, the Bayes factor reduces to

$$B_{1,0} = \frac{B(a,b)}{B(y+a, n-y+b)} \prod_{l=1}^{L} \frac{B(x_l + a_l, n_l - x_l + b_l)}{B(a_l, b_l)} ,$$

where – as usual – $B(a,b)$ denotes the Euler's Beta function with parameters a and b. In the case of non-informative Uniform priors, i.e. when $a = b = 1$ and $a_l = b_l = 1$ for $l = 1, \ldots, L$, it is apparent that $B_{1,0}$ simplifies to

$$B_{1,0} = \frac{\prod_{l=1}^{L} B(x_l + 1, n_l - x_l + 1)}{B(y+1, n-y+1)} .$$

We have applied the testing procedures developed in the previous section to the data of EXP2 by considering the areas as the strata (see Table II). At first, by assuming the frequentist paradigm, we have considered the null hypothesis H_0 of missing proportion homogeneity between strata. The null hypothesis H_0 can be rejected since the P-value corresponding to the test statistic R was less than 10^{-16}. Subsequently, by assuming the Bayesian paradigm and non-informative Uniform priors, we have computed the Bayes factor. In turn, the missing proportion homogeneity is not likely, since $B_{1,0}$ was less than 10^{-16}.

6 Discussion and conclusion

The Italian governmental agency for research evaluation ANVUR conducted two experiments for assessing the degree of agreement between bibliometrics and peer review in the Italian research assessment exercises. They were based on stratified random samples of articles, which were classified by bibliometrics and by informed peer review, respectively. Subsequently, measures of agreement between the ratings produced by the two evaluation methods were computed. Unfortunately, ANVUR provided the final results of both experiments after having dropped from the sample the groups of articles for which the bibliometric algorithm did not produce a score. In addition, as to EXP2, ANVUR also dropped a group of articles for which peer review was unavailable. Therefore, ANVUR presented the results of EXP1 and EXP2 without considering the inferential problems arising from the sample reductions and the missing data.

This paper reconsidered in full the two experiments by adopting the same indicator of

agreement – i.e. the weighted Cohen's kappa coefficient – and also the same systems of weights used in EXP1 and EXP2. In view of analyzing the experiments in the correct inferential setting, the design-based estimation of the Cohen's kappa coefficient and of its confidence intervals were developed and adopted for computing the agreement in EXP1 and EXP2. We proposed three ways of defining in a proper way the population Cohen's kappa coefficients to be estimated. In one case, the suggested definition matched the population coefficient estimated by ANVUR. Regrettably, the estimates for the upper bounds of the confidence intervals of the weighted Cohen's kappa coefficients indicate a degree of agreement that can be considered – at most – weak, since they are smaller than 0.40 for the aggregate population and for each scientific area both in EXP1 and in EXP2.

All the point estimates of the Cohen's kappa coefficients were generally lower in EXP2 than in EXP1. This issue was probably due to the adopted systems of ratings – which are based on four categories in EXP1 and on five categories in EXP2. Moreover, the two systems of weights developed by ANVUR are likely to boost the value of the weighted Cohen's kappa coefficients with respect to other – more usual – systems of weights. Hence, the estimates for the upper bounds of the confidence intervals indicate that the "real" degree of agreement between bibliometrics and peer review is likely to be worst than weak in both EXP1 and EXP2.

The unique exception was Area 13 (Economics and Statistics) in EXP1 for which the Cohen's kappa coefficient was estimated to be 54.17. This outlier can be easily explained by considering that, in EXP1, the Area 13 panel substantially modified the protocol of the experiment with respect to the one adopted in the other areas (Baccini and De Nicolao, 2016a). In EXP2, when the protocol was the same for all the areas, the findings for Area 13 were coherent with those of the other areas.

The results of EXP1 and EXP2 cannot be easily extrapolated. Indeed, the Cohen's kappa coefficients (3), (5) and (7) measure the agreement for the population of articles for which the bibliometric algorithm reached a definite rating. Hence, they cannot instead considered as an agreement measure for the whole set of journal articles produced by Italian scientists in the considered areas. In addition, the articles submitted to the research assessment exercise – from which the sample was obtained – cannot be recognized as representative of the whole set of the journal articles produced in Italy. Indeed, the submitted articles are not randomly selected by researchers, but in view of maximizing the research evaluation results. It is at once apparent that this issue produced a "biased" population of submitted articles.

As to EXP2, a further problem arose for the presence of missing values originated by the refusal of some peer reviewers to referee articles of the sample. These missing values occurred in different proportion in the strata. Thus, under a model-based approach, we have introduced techniques for testing the missing proportion homogeneity between the strata of a population – both under the frequentist and the Bayesian paradigms. Results of the testing procedures show that the missing proportion homogeneity between the scientific areas should be rejected. As a consequence, the sampling selection adopted by ANVUR for EXP2 cannot be considered as "representative" – as claimed by ANVUR – of the population of articles submitted to the research assessment. Indeed, "representativeness" could be solely guaranteed if the missing proportions in the strata induced by article elimination were homogeneous. In the case of EXP2, the results cannot be considered valid for the population of journal articles submitted to the research assessment, nor for the population of journal articles produced in Italy in the considered years.

The findings of this paper are relevant from three points of view. As to the Italian research assessments exercises, they demonstrate that the results of the experiments cannot be considered at all as validating the use of the adopted dual method of evaluation. At the current state of knowledge, it cannot be excluded that the use of a dual method of evaluation introduced uncontrollable major biases in the final results of the assessment. Since the evidence drawn from data in the official research reports shows that peer reviewers' scores were – on average – lower than bibliometric ones, the aggregate results for research fields, departments and universities might be affected by the proportion of research outputs evaluated by the two different tecniques: the higher the proportion of research outputs evaluated by peer review, the lower the aggregate score. As to the scientometric literature, our results conclusively show that the scientific papers – using data produced by ANVUR in the context of the Italian research assessment exercises – must be treated with extreme caution, since they contain major biases due to the use of dual method of evaluation. Unfortunately, these data are actually adopted by the Italian government for distributing fund to universities and research institutions. Finally, as to the academic discussion about the agreement between peer-review and bibliometrics, this paper documented that the two experiments conducted by ANVUR do not bring a valid contribution to the knowledge on that topic, since they have been designed and implemented in a largely unsatisfactory manner.

References

Abramo, G. and D'Angelo, C.A. (2016). Refrain from adopting the combination of citation and journal metrics to grade publications, as used in the Italian national research assessment exercise (VQR 2011–2014). *Scientometrics*, 109(3), 2053-2065.

Abramo, G. and D'Angelo, C.A. (2017). On tit for tat: Franceschini and Maisano versus ANVUR regarding the Italian research assessment exercise VQR 2011–2014. *Journal of Informetrics*, 11(3), 783-787.

Albert, J. (2009). *Bayesian Computation with R* (2nd ed.). New York: Springer.

Alfò, M., Benedetto, S., Malgarini, M. and Scipione, S. (2017). On the use of bibliometric information for assessing articles quality: an analysis based on the third Italian research evaluation exercise. Presentation at the STI 2017, September 6-8, 2017, Paris.

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. and Sileoni, S. (2015). Evaluating scientific research in Italy: The 2004–10 research evaluation exercise. *Research Evaluation*, 24(3), 242-255.

ANVUR (2013). *Valutazione della qualità della ricerca 2004-2010. Rapporto finale.*

ANVUR (2017). *Valutazione della qualità della ricerca 2011-2014. Rapporto finale.*

Baccini, A. and De Nicolao, G. (2016a). Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise. *Scientometrics*, 108(3), 1651-1671.

Baccini, A. and De Nicolao, G. (2016b). Reply to the comment of Bertocchi et al. *Scientometrics*, 108(3), 1675-1684.

Baccini, A. and De Nicolao, G. (2017a). A letter on Ancaiani et al. 'Evaluating scientific research in Italy: the 2004-10 research evaluation exercise'. *Research Evaluation*, 26(4), 353-357.

Baccini, A. and De Nicolao, G. (2017b). 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).

Benedetto, S., Cicero, T., Malgarini, M. and Nappi, C.S. (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.

Berry, K.J., Johnston, J.E. and Mielke, P.W. (2018). *The Measurement of Association.* Switzerland: Springer Nature.

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C.A. and Peracchi, F. (2013). Bibliometric evaluation vs. informed peer review: Evidence from Italy. Unpublished manuscript. Naples: CSEF Working Papers.

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37-46.

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4), 213-220.

De Raadt, A., Warrens M.J., Bosker, R.J. and Kiers, H.A.L. (2019.) Kappa coefficients for missing data. *Educational and Psychological Measurement*, 79(3), 558-576.

Demnati, A. and Rao, J.N.K. (2004). Linearization variance estimators for survey data (with discussion). *Survey Methodology*, 30(1), 17-34.

Fagerland, M.W., Lydersen, S. and Laake, P. (2017). *Statistical Analysis of Contingency Tables*. Boca Raton: CRC Press.

Fleiss, J.L., Levin, B. and Paik, M.C. (2003). *Statistical Methods for Rates and Proportions* (3rd ed.). Hoboken: Wiley.

Franceschini, F. and Maisano, D. (2017). Critical remarks on the Italian research assessment exercise VQR 2011–2014. *Journal of Informetrics*, 11(2), 337-357.

Goukrager, S. (2012). *Objectivity*. Oxford: Oxford University Press.

Gwet, K.L. (2014). *Handbook of Inter-rater Reliability: the Definitive Guide to Measuring the Extent of Agreement among Multiple Raters*. Gaithersburg: Advanced Analytics.

Johnson, N., Kemp, A. and Kotz, S. (2005). *Univariate Discrete Distributions* (3rd ed.). New York: Wiley.

Kass, R.E. and Raftery, A.E. (1995). Bayes factor and model uncertainty. *Journal of the American Statistical Association*, 90, 773-795.

Kulczycki, E., Korzeń, M. and Korytkowski, P. (2017). Toward an excellence-based research funding system: Evidence from Poland. *Journal of Informetrics*, 11(1), 282-298.

Lehmann, E.L. and Romano J.P. (2005). *Testing Statistical Hypotheses* (3rd ed.). New York: Springer.

Pride, D. and Knoth, P. (2018). Peer review and citation data in predicting university rankings, a large-scale analysis. in: Méndez E., Crestani F., Ribeiro C., David G., Lopes J. (eds.) *Digital Libraries for Open Knowledge*. TPDL 2018. *Lecture Notes in Computer Science*, 11057. Cham: Springer.

Quatember, A. (2015). *Pseudo-Populations: A Basic Concept in Statistical Surveys*. New York: Springer.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Development Core Team, R Foundation for Statistical Computing, Vienna, Austria.

Sheskin, D.J. (2003). *Handbook of Parametric and Nonparametric Statistical Procedures* (3rd ed.). London: Chapman & Hall.

Strijbos, J.W., Martens, R.L. Prins, F.J. and Jochems, W.M.G. (2006). Content analysis: What are they talking about? *Computers & Education*, 46(1), 29-48.

Strijbos, J.W. and Stahl, G. (2007). Methodological issues in developing a multi-dimensional coding procedure for small-group chat communication. *Learning and Instruction*, 17(4), 394-404.

Thompson, M.E. (1997). *Theory of Sample Surveys.* London: Chapman & Hall.

Uebersax, J.S. (1987). Diversity of decision-making models and the measurement of inter-rater agreement. *Psychological Bulletin*, 101(1), 140-146.

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Kain, R., Kerridge, S., Thelwall, M., Tinkler, J., Viney, I., Wouters, P., Hill, J. and Johnson, B. (2015). *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management.* London: Sage.

Wolfram Research, Inc. (2014) *Mathematica*, Version 10.0, Champaign, Illinois.