Economica

Economica (2021)
doi:10.1111/ecca.12373

Have you Read This? An Empirical Comparison of the British REF Peer Review and the Italian VQR Bibliometric Algorithm

By Daniele Checchi*, Alberto Ciolfi†, Gianni De Fraja‡, Irene Mazzotta† and Stefano Verzillo§

*ANVUR and University of Milan †ANVUR ‡University of Nottingham, University of Rome and CEPR §European Commission JRC

Final version received 18 February 2021.

This paper determines the assessment of publications submitted to the UK research evaluation carried out in 2014, the REF, which would have resulted if they had been assessed with the bibliometric algorithm used by the Italian evaluation agency, ANVUR, for its evaluation of the research of Italian universities. We find extremely high correlations between the two assessment approaches.

Introduction

In the week before Christmas 2014, university PR offices and common rooms up and down the country were abuzz with discussions and dissections of the freshly published results of the 2014 'Research Excellence Framework' (REF), the official evaluation of all the research conducted by UK academic institutions in the six-year period 2008–13.

This peer-review-based evaluation was the last in a series of such exercises, which have taken place at approximately regular intervals after the initial dummy run held in 1986. It was undertaken jointly by the four UK higher education funding bodies.[1] Its raison d'être was twofold: on the one hand, to ensure accountability for taxpayers' investment in academic research and persuading the general public of its benefits, on the other hand to form the basis for the selective allocation to institutions of the annual 'block' budget for research. The funds allocated on the basis of the results of the REF are around one-quarter of all the funds transferred from the taxpayer to higher education institutions.

Following the previous exercise, held in 2008, the funding agency ran a pilot study with a view to replace peer review, considered very expensive, with an evaluation based on a bibliometric algorithm, but concluded that 'bibliometrics are not sufficiently robust at this stage to be used formulaically or to replace expert review in the REF' (HEFCE 2009). So the 2014 exercise continued to rely on peer evaluation of academic output, although the assessors could choose to use citation information to inform their expert review. The estimated overall cost of the 2014 exercise is approximately £246 million (Farla and Simmonds 2015), comparable to the annual budget of a medium-sized university, and equivalent to £4000 per academic assessed. The next exercise, planned for 2021, but delayed by the Covid-19 pandemic, will also be conducted via peer review, partly because of UK academia's continued opposition to an increased role for mechanical methods of evaluation of research output, even when several other countries do adopt a bibliometric evaluation, as highlighted in the survey of Wang *et al.* (2014). To the extent that considerable cost saving could be achieved by a bibliometric approach, it is not surprising that the literature has addressed the question of the closeness between a peer review and a bibliometric approach. Thus Bertocchi *et al.* (2015) report on the

working method of the economics and management assessment panel in the Italian 2004–10 assessment, which randomly selected some of the journal articles assigned to bibliometric evaluation also to be peer reviewed, precisely to assess the correspondence between the two methods (see also the comment of Baccini and De Nicolao (2016) on the analysis, and the authors' reply, Bertocchi *et al.* (2016)). Mryglod *et al.* (2015) assess the correlation between the score and rank obtained by each institution with the corresponding 'departmental h-index' (Hirsch 2010). The latter paper examines a broader range of research areas than Bertocchi *et al.* (2015), and reports good correlations in the various subject areas, between 0.36 and 0.89. However, it uses a different set of articles from those evaluated by the REF panels, and indeed it includes articles written by academics who were not submitted as part of the group evaluated by the relevant REF panel. In the same vein, Harzing (2018) has shown that ranking UK departments according to the 'departmental h-index' correlates to the REF power ranking at 0.97.

In this paper, we assess the journal articles that were submitted to the UK REF and are included in the Scopus database, with the bibliometric criteria used by the Italian evaluation agency (ANVUR) to assess the outputs, published from 2011 to 2014, submitted for the 2015 Italian evaluation exercise, labelled VQR (eValuation of the Quality of the Research), in the STEM research areas. There are two important differences between this paper and the literature mentioned above. First, we consider *all* the research areas, and, second, we assess only journal articles submitted to the REF, and hence, at least in principle, we compare the two approaches, bibliometric and peer review, on the basis of the same set of research outputs.

We stress at the outset an important limitation of the exercise, which makes its contribution more a template for more thorough analysis than policy advice in its own right: books and book chapters, which constitute an important form of output in some research areas, cannot be assessed by the Italian ANVUR algorithm. There are also several other specific differences between the two evaluations (illustrated in Table 3 below). We did not make any adjustment to the algorithm to account for these. Such adjustments would have an *ad hoc* nature, and one criterion of choice among possible alternative adjustment methods would inevitably be whether or not they improve the correlation between the rankings; as such they would bias our exercise. Even then, we find a remarkable correspondence between the methods: in the 18 REF research areas where at least 75% of the outputs submitted to the REF could be evaluated bibliometrically, the average of the correlation in each REF research area between the average quality of departments in the REF peer review score and the corresponding measure calculated with the ANVUR algorithm is 0.81, and the average rank correlation is 0.76; for the full sample, the figures are 0.63 and 0.60.

Correlation is very much higher for other measures of departmental research quality, which consider the *size* of the unit as well as its average quality; of particular interest to policymakers is the correlation in the funding that would be attributed by the two methods, which stands at 0.995 when the units of assessment with at least 75% of the outputs could be evaluated bibliometrically, and at 0.986 for the whole sample. Even when stacking the deck against the comparison by applying it without making it any allowance for the type of outputs submitted, we show that had the annual funding to institutions been allocated following the ANVUR assessment methods, the outcome would have differed relatively little. Because discrepancies in the various units of assessment tend to cancel each other out, the summary statistic of the correlation in the institutional funding is even more striking: if the output submitted had been evaluated

with the bibliometric algorithm used in the Italian VQR, with peer review assessing the rest of the institutional submission, then the correlation between the actual funding assigned to each institution and the funding it would have received if calculated with the VQR score would have exceeded 0.9997, and hence the difference in funding would have been minuscule. Moreover, much of the difference in funding is due to discrepancies in the institutions with very low funding, which are often specialists institutions, such as art or music schools, whose output is not readily amenable to bibliometric assessment. Of course, uncovering correlations is not establishing causality; we cannot know whether these high correlation scores are due to the fact that the British REF reviewers assess each paper independently, ignoring the bibliometric information about it that they have or can easily access, or because they are in fact, consciously or unconsciously, influenced by the reputation of the journal where the paper appears. To some extent, it does not matter; the paper shows that for whatever reason, a funding allocation very similar to that determined by the 2014 REF peer review could have been obtained in a much less costly manner. Whether our analysis should influence the manner in which future assessments should be carried out therefore requires careful weighing up of the cost savings with the possible downside of this approach that advance knowledge of the assessment mechanism may itself influence the researchers' behaviour in unintended ways, in terms of both which journals to aim for, and even which avenues of research to pursue.

A further result with potential policy consequences is the increase in correlation between peer-review and bibliometric evaluation as the size of the unit evaluated. This confirms the perception held by many (e.g. Harzing 2018) that while they are useful for large sets of researchers, bibliometric algorithms are not suitable for the evaluation of very small groups, let alone individual academics, for example for the purpose of appointment, tenure or promotion decisions.

We close the paper with a simple attempt to uncover association between the closeness of the measure and other institutional variables. We find very little systematic variation: only two variables appear to explain some of the difference in the scores of the two assessment methods. One is the size of the submission, with larger units of assessment appearing to have been slightly penalized by the REF peer review relative to the bibliometric VQR algorithm. The other is the number of units in the institution as a whole: universities with many units of assessment perform a little better with the REF than they would have done with the VQR bibliometric algorithm.

This paper is organized as follows. In Section I we describe the REF evaluation, and in Section II we present the bibliometric algorithm adopted in the Italian VQR. Section III describes the data used to evaluate the REF journal articles, and Section IV reports the results. A brief conclusion ends the paper in Section V.

I. The 2014 Research Excellence Framework

The 2014 REF exercise evaluated the research conducted by 52,000 academic researchers associated to almost 2000 units of assessment (UoAs) in 154 UK higher education institutions. The assessment was carried out by 36 expert panels, one in each area of research (the full list is in Table 5 below), in turn grouped into four 'main panels', corresponding to very broad disciplinary areas: medicine and biology, the other sciences and engineering, the social sciences, and the arts and humanities. The 36 panels comprised over 1000 assessors in total, three-quarters of them academic, the rest non-academic 'users' of the research. The grouping of the disciplines differs in the two

exercises that we consider, the VQR and the REF. It may therefore be useful to fix terminology for the rest of the paper. We denote as 'subject areas' the 350 subject categories in Scopus—this is the finest available classification of topics. In the formal analysis we index the subject areas with h. We then denote as 'VQR research areas' and 'REF research areas' the groups of subject areas that were assessed by the 16 VQR individual panels (known as GEV *Gruppi Esperti Valutatori*) and the 36 REF panels. The correspondence between Scopus subject areas, VQR research areas and REF research areas is close, but by no means perfect; it is in general not possible to map a given journal to a given REF research area, in view of the fact that institutions choose to which panel a given academic is submitted, and the panels may, but are not obliged to, cross-refer a given paper to a different panel (see note 3 below).

The REF panels assess submissions, not individuals or papers. They do so by assessing three main dimensions of an institution's activity:

i. Individual research outputs consisting, for each member of staff submitted, of four outputs published in the reference period 2008–13; outputs can be submitted by an institution as long as the author is employed by that institution on the REF census date, 31 October 2013, irrespective of where the author was when the paper was written or published. The expert panels assessed the output component of each submission, carrying out peer-review evaluations of the 'reach and significance' of each output submitted.
ii. The research environment, as described by each institution in a written submission outlining the achievements of the academics submitted, together with data on research grant income and PhD completions.
iii. The impact of research on the wider society, in terms of knowledge transfer and/or public engagement. Impact is assessed by considering written 'case studies', one for every eight academics submitted, accompanied by supporting evidence that shows how the research of the department has brought benefits *outside of academia*, through, for example, influence on government policy or industry practice. Unlike output, impact is attributed to the institution where the research was carried out, irrespective of which institution is currently employing the researcher responsible for it at the census date. The measures of environment and impact have no exact correspondence in the Italian VQR, and cannot obviously be the object of a bibliometric approach, so we limit our comparison to the output component of the REF.

Having announced the assessment criteria well in advance, the panels determined, on the basis of a peer review of each output submitted, the percentage for each of the three dimensions of the activities of each submission to be assigned to the five quality categories, ranging from the best, 4 stars (or 4*), 'quality that is world-leading in terms of originality, significance and rigour', to the worst, 'Unclassified', 'quality that falls below the standard of nationally recognized work' . We relabel the latter as 0 stars, for consistency with the other categories. On Thursday 18 December 2014, the panels' assessments of each dimension of activity of every institution was made public, together with the aggregate profile, obtained as a weighted average of the outputs, environment and impact components, with the weights 0.65, 0.15, 0.2.[2]

The unit of assessment is the group of researchers submitted to a given national panel; there was no requirement that all the academics submitted to the unit should be all part of an institutional group, such as a department, a school or an institute. Though obviously this was the case for many submissions, there were also many examples of

members of one department being submitted as part of a different unit of assessment from their colleagues. To lighten the exposition, we refer to a department or unit for the group of academics that an institution submitted for assessment as a specific UoA to the appropriate REF panel, but it must be kept in mind that, for example, health economists, behavioural economists, econometricians, political economists, development economists, all working in their economics department, were submitted to the 'Public Health', 'Psychology', 'Mathematical Sciences', 'Politics and International Studies', 'Anthropology and Development Studies' REF panels, respectively. And indeed, many institutions submitted the entire department of economics to the 'Business and Management Studies' panel.[3]

The decisions regarding submissions were taken usually at institutional level, often for tactical reasons, with the attempt to improve the result, and usually had no consequences in the day-to-day lives of the academics or the departments involved. In addition, there was no obligation either to submit all departments for evaluation, or to submit all the academic members of each department submitted. In the event, different institutions took different approaches to the decision whether or not to submit a researcher at all, some leaving out weaker researchers, other including every academic on the payroll. The tactical aspects of the submission strategies, which were heavily influenced by the opinion of how the various panels would judge the quality of the research output, cast strong doubts on the possibility of extending to all disciplines the approach of drawing on departmental information to map the outcome of the REF taken by Mryglod *et al*. (2015) and Harzing (2018) for some of the REF research areas.[4]

Unlike in the Italian assessment, the results are not summarized in a single score that would immediately determine a ranking of institutions. Commentators and the public have therefore stepped in, variously aggregating the profiles into single numbers so as to draw ranking of units of assessment and institutions in national league tables. The most commonly used are the grade point average, GPA, and the research power, RP (Forster 2015). GPA is calculated as a weighted average of the scores, with the proportion in each category as weight: the GPA of UoA i in institution k is calculated simply as

$$GPA_{ik}^{\text{REF}} = \sum_{s=0}^{4} \pi_{ik}^{s} s, \tag{1}$$

where π_{ik}^{s} is the proportion of the activity of UoA i in institution k that was assessed to be of s-star quality. Table 1 reports the summary statistics for GPA in the three components

TABLE 1
SUMMARY STATISTICS AND CROSS-CORRELATIONS OF REF PERFORMANCE BY COMPONENT

	GPA	GPA Outputs	GPA Environment	GPA Impact	Mean	S.D.
GPA	1				2.82	0.433
GPA Outputs	0.93***	1			2.76	0.369
GPA Environment	0.883***	0.71***	1		2.88	0.751
GPA Impact	0.826***	0.578***	0.726***	1	2.98	0.689

Notes
The final sample comprises 1828 units of assessment submitted to REF2014. For explanation of the components, see the text.
*** denotes significance at the 1% level.

and in aggregate. It shows that the correlation between the three components is high, but not so much as to make it meaningless to assess the three components separately.

The other measure widely used to rank departments is research power (RP), which again has no official status. It is simply the product of GPA and the number of staff submitted:

$$RP_{ik}^{\mathrm{REF}} = n_{ik} \times \sum_{s=0}^{4} \pi_{ik}^{s} s, \tag{2}$$

where n_{ik} denotes the number of full-time equivalent researchers submitted by institution k to panel i. *RP* captures the idea that a tiny department of world class quality may be less important, from the viewpoint of research, than a very large department where there are also some academics not as outstanding, who therefore lower the departmental average measured by GPA. Thus the RP measure takes into account the size of the unit of assessment. There is an obvious trade-off between the two: excluding a relatively weak member of staff would definitely increase GPA and reduce research power.

While less prominent in the media, the government, by the very fact of basing the research funding allocations on the results of the REF, does in practice determine a further single measure, which can be used to rank departments within units of assessments, and subsequently aggregated to institutions. This is the funding score, FS, which is part of the formula used to calculate how to allocate the overall 'quality-related' funding made available to the sector in each year. Unlike the funds distributed by the research councils that are strictly linked to specific projects, universities are free to spend this funding as they wish, with no link to projects or even disciplines.[5]

When designing the funding formula, the government intended to provide incentives towards high-quality research, so it gave high weight to 4* output, specifically four times higher than the weight given to 3* output, and *no weight* to output judged less than 3*.[6] With the above notation, the amount of an institution's funding in year t until the following evaluation exercise attributable to UoA i is given by

$$FS_{ikt}^{\mathrm{REF}} = \Phi_t \times \Gamma_i \times (4\pi_{ik}^{4} + \pi_{ik}^{3}) \times n_{ik}, \tag{3}$$

where Φ_t is the coefficient (in the jargon the 'QR unit funding' in year t), which is determined at the beginning of each year, depending on the overall public funding allocated by the government to the university sector. Finally, Γ_i is a research area specific weight which takes value 1.6 for STEM subjects (UoAs 1–15), value 1.3 for intermediate cost research areas (UoAs 16, 17, 26, 34 and 35, which include geography, architecture, sport sciences, design, music), and value 1 for all other research areas.

Table 2 reports the correlation between these measures, indicating that the size-based ones, RP_{ik} and FS_{ikt} (note that Γ_i and Φ_t are constant in each REF research area, and so do not affect the correlation with RP_{ik}), are fairly close to each other, but rather different from GPA_{ik}, which measures the average departmental quality. The correlation between the number of academics submitted, n_{ik}, and the GPA score, GPA_{ik}, is 0.433, indicating that the low correlation between GPA and RP may be due to institutions pursuing different strategies, some preferring prestige, and thus selecting only their best performers, others pursuing the funding associated with larger submissions.

TABLE 2
CORRELATION BETWEEN POSSIBLE MEASURES OF PERFORMANCE

	GPA	Research Power	Funding Score	Mean	S.D.
GPA	1			2.82	0.433
Research Power	0.377***	1		79.62	93.11
Funding Score	0.508***	0.978***	1	38.19	50.96

Notes
Sample: 1828 units of assessment submitted to REF2014. For explanation of the measures, see the text.
*** denotes significance at the 1% level.

The main aim of this paper is to determine the similarity between the two methods of assessment, the REF peer review and the Italian VQR bibliometric measurement. To do so, for each journal article submitted to the REF by every UK institution, we calculate the quality scores that each of its outputs would have obtained if the REF assessment of the output component of the research activity had been carried out using the algorithm that was used by the Italian bibliometric panels to assess the quality of the research of Italian institutions in the 2011–14 period.

We stress that we do *not* attempt to perform a comparison between Italian and British institutions. For example, we do not attempt to compare whether the research in the biology department at the University of Nottingham is better or worse than that carried out in the corresponding department at the University of Rome Tor Vergata. The reason is that, had the British institutions been assessed by the Italian VQR, they would have submitted an altogether different set of outputs, given the many differences between the sets of rules used in the two exercises, illustrated in Table 3.

Differences between the results of the two assessment methods could spring from two sources. On the one hand, there could be structural differences between the methods, which would be the case if a substantial fraction of the highly cited papers published in prestigious journals were, rightly or wrongly, considered to be of poor quality by the peer reviewers, or vice versa, if peer review assessed as being of top quality many papers published in obscure journals with low citation counts. On the other hand, there might be systematic differences in the submission strategies of different institutions: for example, large institutions may be able to devote more resources to assess internally the quality of each output submitted, while smaller ones have to rely on a bibliometric algorithm to select the papers and the academics to submit for evaluation. Of course, a similarity between the VQR bibliometric and the REF peer review assessment could emerge if they did *in general* yield different results, but in the specific case of the 2014 REF, these various factors cancelled each other out. Thus the nature of our paper can be only suggestive, even though, compared to some of the existing literature, it covers the whole of the research carried out in the UK.

II. THE VQR BIBLIOMETRIC ALGORITHM

The Italian assessment exercise required outputs to be classified by the VQR bibliometric algorithm only for the STEM subjects. The panels in other subjects could use some bibliometric information, as the economics panel did, or peer review. All panels had to use peer review for outputs for which this information was not available, such as books,

TABLE 3
DIFFERENCES BETWEEN THE VQR (ITALY) AND THE REF (UK)

	REF	VQR
All departments/units evaluated	No	Yes
All researchers submitted	No	Yes
Portability of output	Yes	Yes
Weight of output in assessment	65%	80%
Period of evaluation	2008–13 (6 years)	2011–14 (4 years)
Census date	31 October 2013	30 November 2014
Number of outputs per person	4	2
Expert panel	Yes	Yes
Peer review	Yes	Depending on VQR research area[a]
Bibliometric indicators	Available: use at the discretion of the panel	Must be used for STEM research areas
Peer review by	Panel members or other panels	Panel members and external reviewers
Overall funding to research area	Depending on evaluation	Predetermined[b]
Funding attributed to	Institutions only	Both institutions and departments[c]
Entity assessed	Department/unit	Individual output

Notes
Summary comparison between the VQR and the REF—see the text for more details. Information obtained from www.ref.ac.uk/2014 (REF) and www.anvur.it/attivita/vqr/vqr-2011-2014 (VQR).
[a]The assessment is bibliometric for the research areas listed in Table 4 (mostly STEM areas), and based on peer review for the others (arts, humanities and social sciences, including economics).
[b]The amount allocated to all the submissions in a given VQR research area is independent of the evaluations given by the VQR panel to the institutions in that research area.
[c]The round of annual funding is allocated to institutions, but a subsequent law awarded a numbers of posts directly to departments, partly on the basis of their VQR score.

contributions to books, or papers in outlets not included in Scopus. We next describe the VQR algorithm. It identifies a paper by four parameters:

i. the year of publication, $t = 1, \ldots, T$;
ii. the subject area, indexed by h;
iii. the number of citations at the census date, appropriately normalized as we explain below;
iv. the impact metrics of the journal where it was published, also normalized.

Because different research areas may have very diverse patterns of citations, the last two parameters are normalized by their position in an appropriate distribution. In detail, to calculate (iii), the VQR algorithm computes the distribution of the citations obtained by all the articles published in subject area h in year t; let this be denoted by $\Psi^C_{ht}(n) \in [0, 1]$. That is, $\Psi^C_{ht}(n) \in [0, 1]$ is the proportion of papers published in subject area h in year t that have obtained n citations or fewer. Similarly for (iv): the relevant measure for journals is the journal impact metric—if a given journal has impact metric x in year t, the algorithm

gives that journal in year t a value $\Psi^J_{ht}(x) \in [0, 1]$ equal to the proportion of journals included in the Scopus database as pertaining to subject area h that, in year t, had impact metric at most x.

For us to calculate the above four parameters characterizing an output submitted to the REF, we need to know the number of citations that it received and the impact factor of the journal where it was published at a given date, and the world distribution of citations and impact metrics for the year in which it was published. The bibliometric information (the number of citations and the SCImago Journal Rank) on 1 January 2015, was obtained from Scopus for each of the papers submitted to the REF. We then used Scopus data on the worldwide distributions of citations and impact metrics for each subject area, to determine the position of each paper and each journal in the worldwide citation count ranking for papers and journals in the subject area.

Having constructed the dataset collecting the four parameters listed at the start of the section, we proceed to the next step of the procedure.[7] The algorithm divides the unit square $[0, 1]^2 \subseteq \mathbb{R}^2$ into five subsets, as shown in Figure 1, by four parallel downward sloping straight lines, in such a way that area I is 0.1, areas II and III are both 0.2, area IV is 0.3, and area V is 0.2. Because we have normalized citation numbers and impact metrics with their rankings, the proportion of the world papers with coordinates that place them in each of the regions is equal to the region's area. Simple computations determine the boundary lines; these are given by $y = a_{ht} - b_{ht}y$, where a_{ht} is the solution in a, for $A = 0.1, 0.3, 0.5, 0.8$, of

$$1 - \max\left\{0, \frac{a-1}{b_{ht}}\right\} - \int_{\min\{1, a/b_{ht}\}}^{\max\{0,(a-1)/b_{ht}\}} (a - b_{ht}x)\mathrm{d}x = A.$$

This solution is given by

$$a_{ht}(A, b_{ht}) = \begin{cases} 1 + \frac{1}{2}b_{ht} - A & \text{if } A \le \frac{1}{2}b_{ht}, \\ 1 - \sqrt{2Ab_{ht}} + b_{ht}(1-x) & \text{if } \frac{1}{2}b_{ht} < A \le 1 - \frac{1}{2}b_{ht}, \\ \sqrt{2b_{ht}(1-A)} & \text{if } A > 1 - \frac{1}{2}b_{ht}. \end{cases} \tag{4}$$

In equation (4), b_{ht} is the slope used to assess outputs in the subject area h in year t; it is chosen subjectively by the panel for each VQR research area, to reflect the trade-off between visibility of an article and prestige of the publishing journal, in their research area, and the manner in which it changes with time. In practice, the slope b_{ht} varied from year to year and from VQR research area to VQR research area, to account for the different citation patterns and the fact that more recent papers have less opportunity to collect citations than equally influential article published five years before, so for more recent papers the impact metric of the journal was given a higher weight. Because of these considerations, the slope of the lines separating the areas in Figure 1 increased in absolute value with the year of publication so as to reduce the importance of citations for younger articles.

A frequent criticism of bibliometric evaluation is that it induces conformism and discourages both heterodox research and emerging journals.[8] It may also distort submission decisions: for example, researchers may try to target journals 'just above the

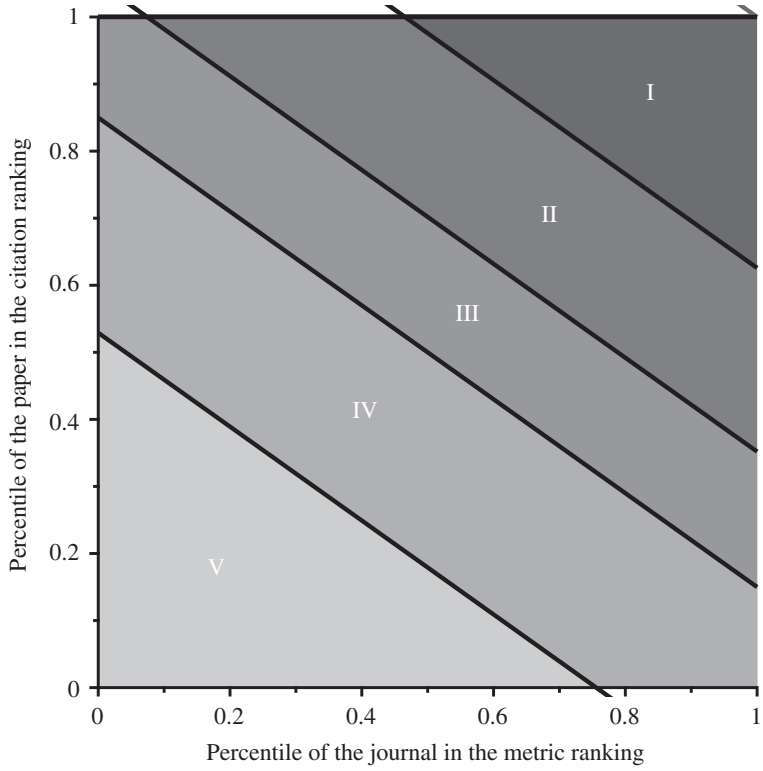

FIGURE 1. Allocations of products to quality classes.

boundary'. Relative to the issuing of an official ranking of journals, the potential for tactical behaviour is lessened by the Italian VQR algorithm, because the ranking of each given publication outlet is not known prior to submission, since its position in the worldwide ranking is endogenously determined at the end of the year of reference. In addition, the status of the journal is only one component of the evaluation; the other is the idiosyncratic success of the paper, measured by the number of citations.

Table 4 reports the slopes that were used in the Italian exercise, and those that we have used to obtain the score for each of the articles that we have assessed. The overlap between the REF and the VQR is such that we could use the VQR slopes only for the years 2011–13. For the other years, rather than arbitrarily setting a pattern of change, we deliberately chose to reduce our degrees of freedom by setting the slopes outside the overlap period to be the same as at its beginning.[9]

The VQR algorithm can now be 'fed' the coordinates of the impact metric ranking and the citation ranking of each paper submitted to the REF, and published in a journal included in the Scopus database to determine a score in {0, 0.1, 0.4, 0.7,1} depending on in which subset of the unit square these coordinates lie. That is, the score assigned to an article published in a journal included in subject area h in year t depends on the number of citations that it received relative to the world distribution of citation for articles published in subject area h in year t, and on the impact metric of the journal where it was published, again relative to the distribution of the impact metrics of journals in subject area h in year t. In detail, consider an article that was in percentile p^C of the world distribution of citation for articles published in subject area h in year t, published in a

TABLE 4
SLOPES OF TRADE-OFFS BETWEEN CITATIONS AND IMPACT FACTOR

	VQR				REF		
VQR research areas	2011	2012	2013	2014	2008–11	2012	2013
Computer Science	1	1.25	1.5	1.75	1	1.25	1.5
Mathematics	Depending on sub-area				1.1	1.4	1.7
Physics	0.4	0.6	0.9	1.5	0.4	0.6	0.9
Chemistry	0.4	0.6	0.8	1.2	0.4	0.6	0.8
Earth Sciences	0.4	0.6	0.9	1.5	0.4	0.6	0.9
Biology	0.4	0.6	0.8	1.2	0.4	0.6	0.8
Medicine	0.4	0.6	0.8	1.2	0.4	0.6	0.8
Agriculture and Veterinary Sciences	0.7	0.9	1.5	2	0.7	0.9	1.5
Architecture	0.6	0.9	1.5	2	0.7	0.9	1.5
Civil Engineering	0.7	0.9	1.5	2	0.7	0.9	1.5
Industrial and Information Engineering	0.4	0.6	0.9	1.5	0.4	0.6	0.9
Psychology	0.4	0.6	1	1.5	0.4	0.6	1

Notes
The table reports the slopes of the lines in Figure 1, in different years, for the VQR research areas that used the bibliometric algorithm. The first four columns report the coefficients used in the VQR, the last three those that we have used to compute the scores of papers submitted to the REF.

journal whose impact metric placed it in percentile p^J of the corresponding world distribution of journals' impact metrics. Then this article's score is given by

$$s_{\text{VQR}} = \begin{cases} 1 & \text{if } p^C \geq a_{ht}(0.1, b_{ht}) - b_{ht}p^J, \\ 0.7 & \text{if } a_{ht}(0.1, b_{ht}) - b_{ht}p^J > p^C \geq a_{ht}(0.3, b_{ht}) - b_{ht}p^J, \\ 0.4 & \text{if } a_{ht}(0.3, b_{ht}) - b_{ht}p^J > p^C \geq a_{ht}(0.5, b_{ht}) - b_{ht}p^J, \\ 0.1 & \text{if } a_{ht}(0.5, b_{ht}) - b_{ht}p^J > p^C \geq a_{ht}(0.8, b_{ht}) - b_{ht}p^J, \\ 0 & \text{if } p^C < a_{ht}(0.8, b_{ht}) - b_{ht}p^J, \end{cases}$$

where, in each row, the dependence of a_{ht} on A and b_{ht} derived in equation (4) is made explicit. In words, an article is considered as: 'excellent' (score 1) if it corresponds to the best 10% in the world joint distribution of citations and journal metric, that is, if its coordinates make it fall in area I in Figure 1); 'good' (score 0.7, area II) if it falls between 10% and 30%; 'fair' (score 0.4, area III) if it falls between 30% and 50%; 'acceptable' (score 0.1, area IV) if it falls between 50% and 80%. The remaining papers are labelled as 'limited', and receive score 0 (area V).

Given that different VQR research areas use different slopes, the unit square depicted in Figure 1 may be divided up differently by the different VQR research areas, so it may happen that a given paper is assessed differently according to the VQR research area to which it is assigned. The assignment of journals to VQR research areas is therefore important to determine the score that an article included in Scopus and submitted to the REF would have received had it been assessed with the VQR bibliometric algorithm. Approximately 70% of the outputs included in Scopus submitted to the REF are published in journals that the VQR had assigned to one VQR research area, including many non-STEM journals. To assign the remaining 30%—almost all outputs published in journals in social sciences arts and humanities—we exploited information on the

frequency of publications in journals of a given Scopus subject area by the academics submitted to a VQR STEM research area.[10] At the end of these steps, we found that around 46% of the outputs submitted to the REF and contained in Scopus were published in journals that are associated to more than one VQR research area. While this may lead to a paper being given different scores according to the different 'slope' coefficients of the different VQR research areas, this happened in only 7068 articles, 5% of those that we assessed. In these cases, we chose the highest evaluation score.[11]

In the final stage of our computation, after each output was assigned to the corresponding class, we calculate the score for each unit of assessment in each institution by aggregating all the scores for each article submitted by that unit. The corresponding score for each institution i evaluated according to the VQR algorithm is given by

$$GPA_{ik}^{\mathrm{VQR}} = 4\pi_{ik}^{1} + 3\pi_{ik}^{0.7} + 2\pi_{ik}^{0.4} + \pi_{ik}^{0.1}, \tag{5}$$

where π_{ik}^{s} is the proportion of the articles of unit of assessment i in institution k to which the algorithm assigned a score $s_{\mathrm{VQR}} = s$, $s = 1, 0.7, 0.4, 0.1$. We perform the comparison at the level of the unit of assessment, not at the 'paper level', because the REF does not have or is not disclosing this valuation. Note, of course, that in equation (5), $\sum_s \pi_{ik}^{s} \leq 1$, but this sum can be strictly less than 1, as some outputs may score zero. In equation (5), we calculate the GPA with the weight vector (4, 3, 2, 1, 0) used in the REF. As a robustness check, we did the same calculation with the VQR weight vector, which was (1, 0.7, 0.4, 0.1, 0). At 0.998, the overall correlation between the measures is very high.

III. The Data

All the outputs submitted to the REF are available from the REF website (www.ref.ac.uk/2014) as Excel files.[12] For each output, the file contains the type of output (journal article, book, working paper, etc.), the institution that submitted the output, and the unit of assessment to which it was submitted, as well as standard bibliographic information such as DOI, publication year, number of co-authors, title, place of publication, and so on. The names of the authors are not included (though of course they are easily obtained), as it is not relevant to the REF, and hence not to our exercise either.[13]

The total number of outputs assessed is 190,962, with 81.09% of the total (154,854) journal articles, the remainder consisting mainly of chapters in books (7.5%) and books (5.4%). There are several other different types, all representing a tiny fraction of the total, such as compositions (0.35%), patents (0.06%), exhibitions (0.65%) or scholarly editions (0.19%).

Scopus returned the required data for 139,847 of the submitted journal articles, the remaining ones having being published in outlets not covered by Scopus. These were books, editorials, notes and the like, and articles in journals not included in Scopus. In addition, a handful of other products could not be evaluated, for various reasons (301 were of a type not considered by the VQR algorithm, such as chapters in books, or monographs included in Scopus, 61 were allocated in the REF published data to an anonymized UoA, and 17 had missing data that made their allocation impossible). The final tally of outputs that we assessed was thus 139,468, nearly two-thirds of the total.

Table 5 presents summary statistics of the output data: as one would expect, the research areas with the highest proportion of outputs that can be assessed using the VQR

TABLE 5
SUMMARY STATISTICS OF THE PAPERS SUBMITTED TO REF2014

Unit of Assessment		Number of institutions	Output in VQR	% of REF submissions	% assessed by REF as 4*	3*	2*	1*	0*
Main panel A		121	48,356	94.44	37	44	17	1	1
Clinical Medicine	1	31	13,400	97.34	39	44	15	1	1
Public Health	2	32	4881	93.26	39	41	17	3	0
Allied Health Professions	3	82	10,358	93.33	31	50	17	1	1
Psychology	4	81	9126	97.04	38	40	19	2	1
Biological Sciences	5	44	8608	98.18	37	46	15	1	1
Agriculture and Veterinary Science	6	29	3919	96.61	35	41	20	3	1
Main panel B		105	44,830	89.11	26	57	15	2	0
Environmental Sciences	7	44	5184	96.53	24	59	15	2	0
Chemistry	8	37	4698	98.47	28	63	9	0	0
Physics	9	41	6446	97.91	28	60	11	1	0
Mathematics	10	53	6994	90.65	29	55	15	1	0
Computer Science	11	89	7651	67.39	26	44	24	5	1
Chemical and Manufacturing Engineering	12	22	4143	95.73	25	57	17	1	0
Electrical Engineering	13	32	4025	96.77	25	62	11	2	0
Civil Engineering	14	14	1384	92.41	24	56	16	3	1
General Engineering	15	62	8679	95.09	26	56	16	2	0
Main panel C		124	36,432	67.61	27	42	26	4	1
Architecture	16	43	3781	66.81	29	40	25	6	0
Geography and Archaeology	17	58	6017	76.32	27	42	26	5	0
Economics and Econometrics	18	28	2600	86.88	30	48	19	2	1
Business and Management Studies	19	98	12,202	89.08	26	43	26	4	1
Law	20	65	5522	30.21	27	46	23	4	0
Politics and International Studies	21	55	4365	60.34	28	40	26	6	0
Social Work and Social Policy	22	62	4784	64.61	27	42	25	5	1
Sociology	23	29	2630	64.90	27	45	26	2	0
Anthropology and Development Studies	24	21	2013	57.68	27	42	26	4	1
Education	25	75	5519	65.43	30	36	26	7	1
Sport Sciences, Leisure and Tourism	26	50	2757	83.9	25	41	27	6	1
Main panel D		138	9850	25.55	30	41	24	4	1
Area Studies	27	22	1724	40.55	28	42	25	5	0
Modern Languages and Linguistics	28	47	4932	27.58	30	42	23	4	1
English Language and Literature	29	86	6923	19.20	33	41	22	4	0
History	30	81	6431	31.27	31	44	23	2	0
Classics	31	22	1386	12.77	34	42	22	2	0
Philosophy	32	39	2173	46.71	31	42	24	3	0
Theology and Religious Studies	33	31	1558	20.54	28	40	27	5	0

TABLE 5
CONTINUED

Unit of Assessment		Number of institutions	Output in VQR	% of REF submissions	% assessed by REF as 4*	3*	2*	1*	0*
Art and Design	34	71	6321	15.57	26	42	25	6	1
Music, Drama and Dance	35	72	4246	16.77	29	39	24	6	2
Media Studies	36	69	3517	35.34	29	38	24	8	1
Total		154	139,468	64.20	19	45	29	5	1

Notes
The columns report the names and numbers of the units of assessment, grouped in their respective main panels, the number of institutions submitted, the percentage of the output submitted that could be assessed with the VQR bibliometric algorithm, and the percentage of the outputs submitted that were assessed by the REF panel as 4*, 3*, 2*, 1* and 0*.

bibliometric algorithm are those in the STEM research areas, for which the VQR algorithm was designed, and those, like economics, where the typical publication outlets are refereed journals.

IV. RESULTS

Our main results are reported in Table 6. The UoAs for the REF are ordered according to the percentages of output that we have been able to assess using the VQR, as in Table 5; the number and letter in parentheses following the name are the REF panel and main panel indicators.

Column (1) of Table 6 reports the correlation between the individual GPA scores calculated for the outputs of the various institutions that submitted to the corresponding UoA using the VQR algorithm (formula (5)), and the scores awarded to these units by the REF expert panel. Column (2) reports the rank correlation between these sets of scores. These two sets of correlations are themselves highly correlated (0.973). All the correlations are positive, and many, especially for the UoAs where a large percentage of the products submitted could be assessed with the bibliometric algorithm of the VQR, are very high; this is true for both the correlations between values and the rank correlations. The GPA measures are averages; they assess the quality of the 'typical' researcher in the unit, and so are independent of the number of academics submitted. When the latter are allowed into the picture, the correlations increase radically, as shown in columns (3) and (4), which report the correlations in RP, and even more so in columns (5) and (6), which report the correlations in the FS measure, the funding attributed to each unit submitted. In column (5), which reports the calculation of the total funding that would accrue to the institution if the scores were calculated using the VQR algorithm, we see extremely high correlations. Their weighted average across REF research areas (with weights that the output submitted to the REF) is 0.989. In more than half the REF research areas, the correlation exceeds 0.99, with the lowest value at 0.913, for 'Music, Drama and Dance'. Even the latter is extremely high, considering that we could assess less than 17% of the outputs in this REF research area. The very high values of the correlations even for REF research areas where relatively few outputs were in Scopus journals (those at the end of Table 6), probably have a very natural explanation.

TABLE 6
CORRELATION IN THE MEASURES AND THE RANKINGS

REF panel	Corr. GPA (1)	Spearman GPA (2)	Corr. RP (3)	Spearman RP (4)	Corr. FS (5)	Spearman FS (6)
Chemistry (8 B)	0.857	0.788	0.987	0.975	0.995	0.993
Biology (5 A)	0.884	0.747	0.989	0.972	0.998	0.993
Physics (9 B)	0.896	0.828	0.992	0.977	0.998	0.993
Medicine (1 A)	0.753	0.811	0.988	0.994	0.999	0.997
Psychology (4 A)	0.847	0.875	0.984	0.963	0.998	0.990
Electrical Engineering (13 B)	0.825	0.808	0.976	0.956	0.993	0.988
Agriculture (6 A)	0.777	0.691	0.977	0.975	0.996	0.993
Environment (7 B)	0.794	0.763	0.980	0.983	0.996	0.991
Chemical Engineering (12 B)	0.690	0.613	0.972	0.943	0.991	0.985
General Engineering (15 B)	0.785	0.780	0.965	0.952	0.994	0.989
Health Professions (3 A)	0.820	0.800	0.979	0.969	0.996	0.991
Public Health (2 A)	0.909	0.761	0.994	0.947	0.999	0.995
Civil Engineering (14 B)	0.832	0.846	0.930	0.951	0.991	0.991
Mathematics (10 B)	0.779	0.680	0.987	0.965	0.998	0.993
Management (19 C)	0.818	0.852	0.985	0.969	0.996	0.996
Economics (18 C)	0.899	0.880	0.987	0.917	0.996	0.973
Sport Sciences (26 C)	0.522	0.467	0.899	0.807	0.985	0.963
Geography (17 C)	0.834	0.777	0.954	0.954	0.994	0.988
Computing (11 B)	0.758	0.665	0.933	0.909	0.989	0.979
Architecture (16 C)	0.624	0.600	0.950	0.859	0.993	0.982
Education (25 C)	0.565	0.575	0.966	0.819	0.996	0.981
Sociology (23 C)	0.542	0.460	0.904	0.933	0.983	0.988
Social Work (22 C)	0.649	0.638	0.907	0.837	0.987	0.980
Politics (21 C)	0.666	0.646	0.957	0.907	0.994	0.982
Anthropology & Development (24 C)	0.308	0.381	0.844	0.836	0.982	0.990
Philosophy (32 D)	0.557	0.521	0.978	0.944	0.988	0.978
Area Studies (27 D)	0.357	0.299	0.890	0.782	0.974	0.928
Media Studies (36 D)	0.443	0.495	0.813	0.788	0.962	0.952
History (30 D)	0.623	0.623	0.967	0.914	0.995	0.984
Law (20 C)	0.612	0.598	0.896	0.861	0.987	0.976
Modern Languages (28 D)	0.001	0.066	0.812	0.715	0.964	0.945
Theology (33 D)	0.400	0.369	0.742	0.686	0.967	0.939
English (29 D)	0.289	0.234	0.868	0.800	0.967	0.958
Music (35 D)	0.136	0.142	0.586	0.487	0.913	0.874
Arts (34 D)	0.211	0.308	0.836	0.664	0.960	0.902
Classics (31 D)	0.345	0.336	0.899	0.684	0.979	0.852

Notes
Comparison between the score and the rank obtained using the VQR algorithm and the actual REF score. GPA, RP and FS are calculated according to equations (1), (2) and (3), respectively. The REF panels are ordered by the proportion of outputs that the VQR bibliometric algorithm can assess (as in Table 5), and the space separates the REF panels where more than 75% of the products could be assessed with the VQR bibliometric algorithm from the rest. The number and letter in parentheses after a UoA's name are the number of UoAs evaluated and the main panel. Pairwise correlations between each pair of correlations are respectively 0.973, 0.778 and 0.903, all three values with $p \leq 0.1$.

It would follow if the quality of the outputs submitted to journals and the quality of the books and other forms of outputs in these REF research areas were correlated; in other words, departments whose members can hit the best journals in the humanities also have members who write the best books. Of course, all our analysis unearths is correlation, and we cannot say anything at all regarding the direction of causality. It may be that high-quality papers are more likely to be submitted to and accepted by good journals because of a good editorial process, and therefore would be judged to be of high quality even in blind peer review; or it might be that the REF reviewers were influenced by the reputation of the journal, and assigned high scores to a paper in view of the prestige of the outlet rather than its intrinsic merit.

The results for the rank correlation are less extreme. Its lower value is likely to be due to the fact that many scores are very tightly bunched, so small measurement errors change little in the absolute scores, but may have large impact in the ranking. Given that the aim of the UK exercise is to assess research, not rank institutions, this is the less relevant of the two correlation measures.

The data reported in Table 6 are also shown in Figure 2, which illustrates the correlations and the rank correlations in the various REF panels according to the three measures that we have considered in this paper. The high correlation in institutional funding shown in column (5) of Table 6 is highlighted by the circles in the left-hand panel of the figure.

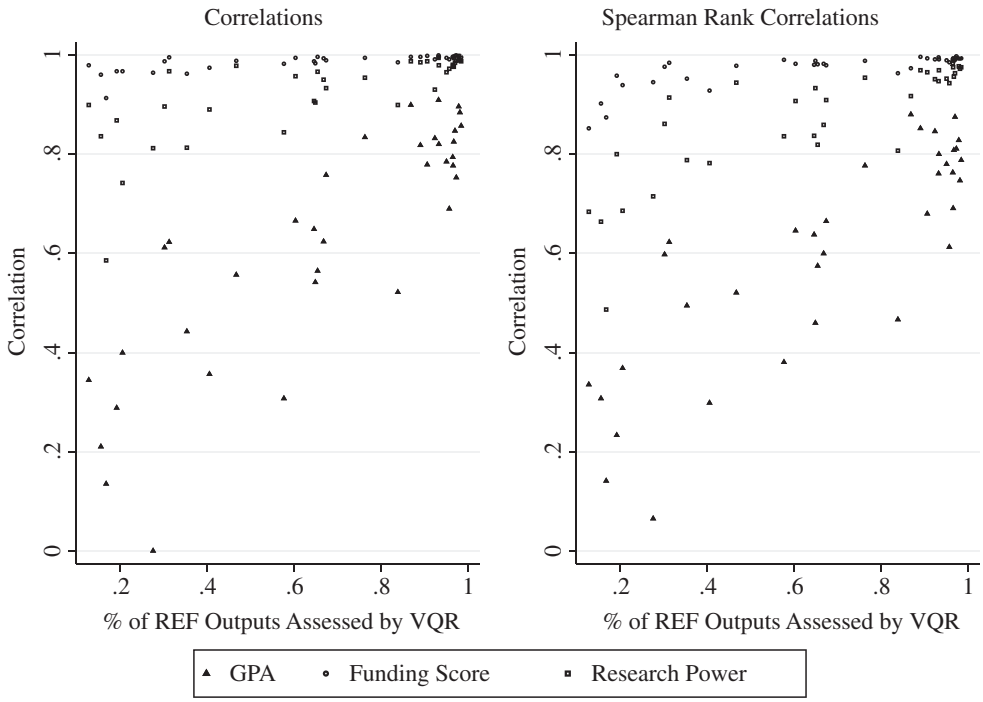

FIGURE 2. Correlations between performance scores.
Notes: The diagrams report the correlations in each REF research area (left-hand panel), and the correlations (right-hand panel) between the score obtained using the VQR bibliometric algorithm and the actual REF scores in the REF2014 assessment. For the three measures considered, see the text: the formal definitions are in equation (1) for the GPA, in equation (2) for the research power, and in equation (3) for the funding score.

We can further explore the link between the scores awarded by the REF peer review and the VQR. While we can compute the score that each individual paper would have obtained had it been assessed with the VQR bibliometric algorithm, the scores attributed to the paper submitted by the REF panel are closely guarded secrets. Therefore in Figure 3 we attempt to disaggregate the scores by reverse engineering them. We do so by plotting against one another the proportions of outputs assigned to the five categories by the two methods. The relatively low correspondence at the panel level depicted in the top row reflects the difference in the overall assessment at the panel level, a result highlighted by De Fraja *et al*. (2019), who show, for example, that the Economics and Econometrics panel had an overall higher standard, and so presumably lower scores in the REF than in the VQR. As long as this applies linearly to all UoAs assessed by the panel, this would not affect the correlation in the departmental scores. The middle and bottom rows are the corresponding figures at the UoAs (the former a binscatter of all the UoAs, the latter only the UoAs larger than the median). These diagrams show stronger correlation, suggesting that once the panel reached an overall standard for all the submissions, the relative ranking of the UoAs within the panel was consistent between the two methods, the REF peer review and the VQR bibliometric algorithm.

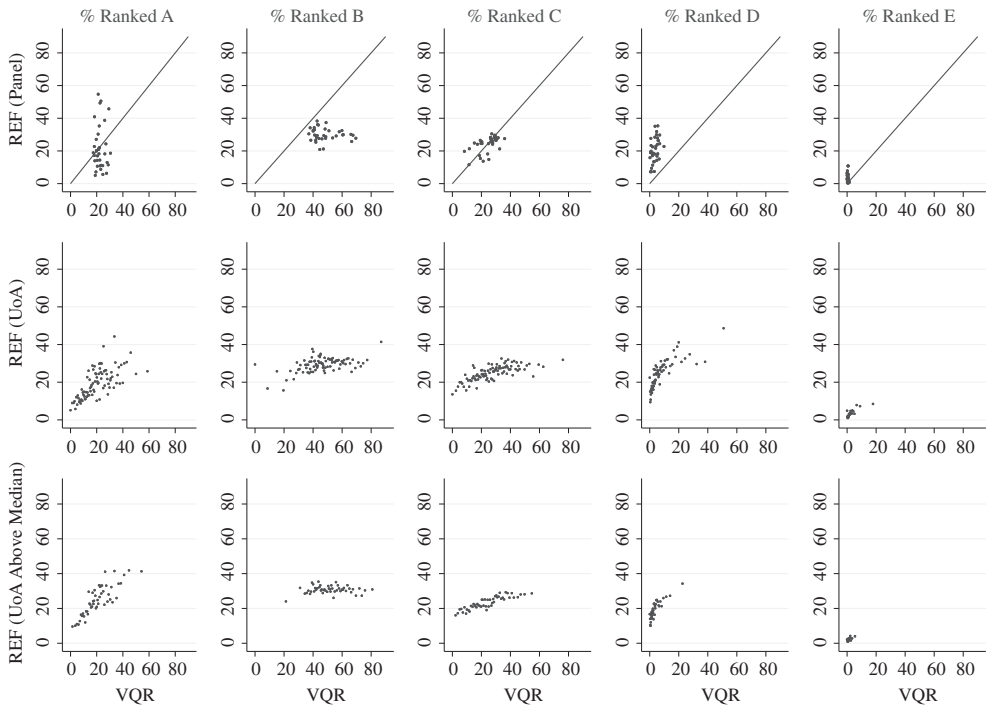

FIGURE 3. The proportion of papers allocated to each category by the two methods.
Notes: Each dot in the first diagram in the top row plots the proportion of the outputs that could be classified as the top category using the VQR bibliometric algorithm submitted to one of the 36 panels against the proportion of the outputs (including those that cannot be allocated with the VQR bibliometric algorithm) assessed by the REF peer review panel to be in the top category (assessed as 4*). Analogously for the other four panels. In the middle row, we report the binscatter plot of the 1828 UoAs: each of the hundred dots represents the average of around 18 UoAs. Because one might expect more divergence for the smaller UoAs, the third row repeats the exercise of the second but taking only the UoAs of above-the-median size.

While we stress once again the highly stylized nature of our computations, it might nevertheless be intriguing to verify, along the lines of Harzing (2018), how the allocation of the governmental funds would have changed if instead of the peer review, the funding agency had assigned funds to universities using the VQR algorithm. This back-of-an-envelope calculation finds justification in the fact that funding is allocated to institutions, *not* departments, so differences in the amount of funding to different units in the same university determined by the two methods may cancel out in the overall institutional funding. Clearly, we change the values of the institutions' scores only for the outputs component of the REF submissions, with everything else—namely the assessment of the environment and of the impact of the research—being held constant. That is, we calculate expression (3) and then sum for all the UoAs that each institution submitted, with two different values of the 'output' performance, one obtained with the VQR assessment and one with the peer review assessment. Equation (3) can be written as

$$FS_{ikt}^{\mathrm{REF}} = \sum_{i \in I_k} n_{ik} \Gamma_i \sum_{s=3,4} 4^{s-3} (0.65 \pi_{ik}^{s,OUT} + 0.15 \pi_{ik}^{s,ENV} + 0.2 \pi_{ik}^{s,IMP}), \tag{6}$$

where $\pi_{ik}^{s,X}$ is the proportion of activity X submitted by unit i in institution k assessed to be of quality s-stars, $s = 3, 4$, with X taking values OUT (output), ENV (environment) and IMP (impact), Γ_i is the cost adjustment parameter taking value 1.6, 1.3 or 1, as explained above, and I_k is the set of units of assessment submitted by institution k. Replacing the REF evaluation of the output with that obtained from the VQR bibliometric algorithm generates an extremely high correlation of the institutions' funding: 0.9997, both when all units of assessment are considered and when only those where at least 75% of the outputs could be assessed with the VQR algorithm.

Lest it be thought that this very high correlation is due to the fact that the environment and the impact components, which together contribute a third to the total, are identical in the two methods of computing an institution's funding, we have calculated the funding that each institution would receive if only the output component were used to determine it. Formally, when we replace the weights (0.65, 0.15, 0.2) in equation (6) with weights (1, 0, 0), the correlation between the institutions' funding is 0.9939, only fractionally lower.

This finding is illustrated dramatically in Figure 4. It plots the annual funding that an institution would receive had the assessment of its outputs being performed with the VQR bibliometric algorithm, against the total funding that it receives every year as a consequence of the performance of its academic units as measured by the peer review mechanism of the REF. The hollow circles in the figure illustrate how close the numbers are, with the exception of some of the smallest institutions, many of which are specialists different from the traditional university.[14] Excluding these institutions increases the correlation still further, to 0.9998. The solid squares are the levels that would result if only output were used to determine funding.

We end the paper by trying to uncover links between any discrepancy in the two methods of evaluating outputs, our calculations using the VQR bibliometric algorithm and the REF peer review evaluation, and observable characteristics of institutions and units of assessment. We are well aware that it is very hard to establish any causal effect, so the results presented in Table 7 should be properly considered simply as a suggestive description. This table reports the estimated coefficients for various specifications of the equation

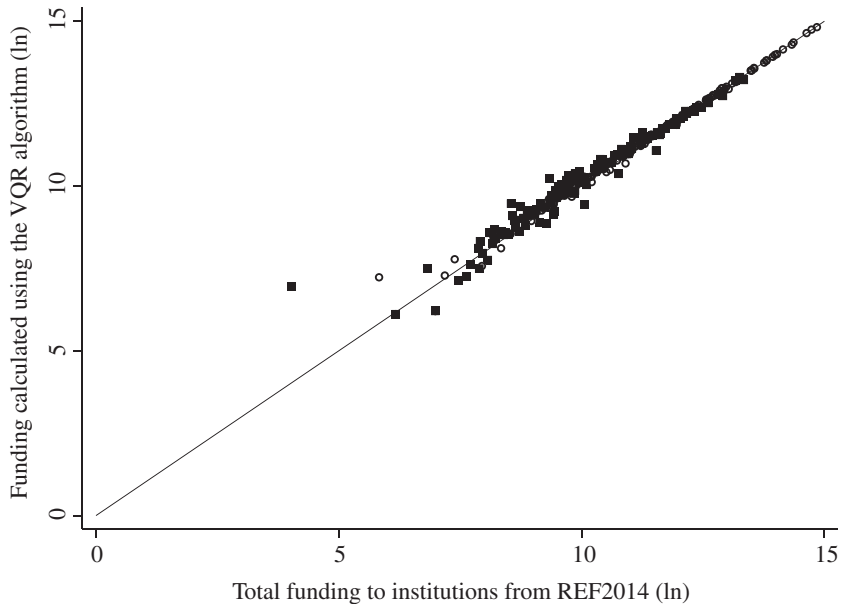

FIGURE 4. Institutional funding with the two assessment mechanisms.
Notes: Institutional annual funding of the UK institutions, allocated according to the REF results (horizontal axis) and the funding that would be determined if outputs had been assessed using the VQR bibliometric algorithm (vertical axis). Axes are in logarithmic scales, and when a point is on the diagonal line, the institution represented by it would receive the same funding under both mechanisms. The hollow circles compute funding using the weights used by the UK funding agency; the solid squares consider the case when all funding is determined by output only.

$$\Delta_{ik} = \beta_0 + \beta_1 n_{ik} + \beta_2 N_k^U + \beta_3 N_{ik}^M + \beta_4 p_{ik} + \beta_5 w_k + \phi_i + \mu_t + \varepsilon_{ik}. \tag{7}$$

In equation (7), $_{ik}$ is the difference in a given measure of research quality or of the corresponding rank between the outcome measured by the VQR algorithm and that assessed by the REF peer review; thus $_{ik}>0$ indicates that the submission to the REF research area i made by university k did better with the VQR algorithm than it was judged to be by the peer reviewers. In the top part of Table 7, we include all the REF research areas. In the bottom part, we restrict the sample to the REF research areas where the percentage of outputs that we were able to assess exceeded 75%.

On the right-hand side of equation (7), we include n_{ik}, the number of academics submitted; this might affect the submission if a larger department might have more resources to devote to preparing the submission (for example, some departments hired an external reviewer to assist them). N_k^U and N_{ik}^M are the numbers of other submitted units in the entire university k and in the same 'main panel' as REF research area i, respectively: the idea here is that if there are many different submissions, it might be easier for an institution to submit academics tactically to different panels with the aim of improving their return.

We include two further variables that De Fraja *et al*. (2019) suggest may be associated with the outcome of the REF evaluation. The first, which varies only at

TABLE 7
DETERMINANTS OF THE DIFFERENCE IN SCORES BETWEEN THE VQR AND THE REF

Dependent variable: (VQR-REF)	GPA		Research Power		Funding Score	
	Full Sample	Restricted Sample	Full Sample	Restricted Sample	Full Sample	Restricted Sample
FTE submitted	0.1035***	0.0639*	0.0003	−0.0002	0.0002	−0.0000
	0.039	0.038	0.001	0.000	0.000	0.000
Other UoAs	−0.1270	0.1411	−0.0106***	−0.0063**	−0.0024**	−0.0012
	0.215	0.264	0.003	0.003	0.001	0.001
Other UoAs in main	−0.0046	0.0898	0.0017	0.0002	−0.0009	0.0002
	0.505	0.679	0.007	0.008	0.002	0.002
Panel member	4.9390**	6.8418**	0.0094	0.0376	−0.0027	0.0152*
	2.158	2.672	0.030	0.030	0.010	0.009
Head's salary	0.0514***	0.0650***	−0.0000	−0.0002	0.0001	0.0001
	0.014	0.019	0.000	0.000	0.000	0.000
Observations	1732	803	1676	801	1731	803
R-squared	0.554	0.616	0.456	0.396	0.407	0.506
FTE submitted	0.0129	0.0006	0.0136	0.0040	−0.0004	−0.0011
	0.017	0.014	0.010	0.006	0.005	0.003
Other UoAs	−0.2832***	−0.0963	−0.1561***	−0.0320	−0.0662**	−0.0137
	0.095	0.095	0.055	0.043	0.026	0.020
Other UoAs in main	−0.2124	−0.0668	0.0129	0.0886	0.0149	0.0371
	0.224	0.245	0.129	0.111	0.062	0.052
Panel member	−0.5952	1.1562	−0.6515	0.4833	−0.1697	0.0569
	0.958	0.963	0.552	0.436	0.263	0.203
Head's salary	0.0007	−0.0055	−0.0012	−0.0026	0.0004	−0.0000
	0.006	0.007	0.004	0.003	0.002	0.001
Observations	1732	803	1732	803	1732	803
R-squared	0.064	0.015	0.083	0.045	0.188	0.151

Notes
Standard errors are robust to heteroscedasticity and clustered at panel level. Determinants of the difference in the result obtained with the VQR bibliometric algorithm and the actual REF score. In the upper part of the table the dependant variable of the OLS regression is the score: the GPA from equation equation (1), the log of the research power from equation (2), and the log of the funding score from equation (3). The restricted sample includes only the UoAs where the VQR algorithm could assess at least 75% of the outputs submitted (those in the lower part of Table 6). The lower part of the table repeats the OLS regression using the rank instead of the score or its log.
***, **, * indicate $p \leq 0.01$, $p \leq 0.05$, $p \leq 0.1$, respectively.

institution level, is w_k, the salary of the head of the institution, usually called Vice-Chancellor, for the year preceding the REF. The second is a dummy p_{ik} indicating that institution k had one of its academics as a panel member for REF research area i. This might be a variable associated with systematic differences as it might be the case that institutions that did have a panel member in the relevant REF research area may have superior insight as to the way in which the assessment will be conducted, and be better able to judge, for example, the opportunity of submitting a certain article with fewer citations or appearing in a less prestigious journal, but with some characteristics that made it more likely to be evaluated highly by the panel.[15] Finally, we include REF research area fixed effects, $_i$, and four dummies to characterize the 'university type' μ_τ, as classified by De Fraja *et al.* (2019), who divide all UK institutions into different types (i.e.

'Russell', '1994 group', etc.) to capture the possibility that those in each group might have different experiences and attitudes to research.

Table 7 suggests that there is very little explanatory power from any of the variables, and in the cases when there is, such as the size of the submissions, the number of other submissions made by the institution, and the presence of a member of the department in the peer review panel, these variables appear to affect only some of the difference in the rankings. Overall differences in scores and rankings between units of assessment in the two exercises, the British REF and the Italian VQR, seem to be due mostly to random non-systematic factors.

V. Concluding Remarks

In this paper, we have performed a simple exercise to compare the outcome of the assessment of the research carried out in British universities in the course of the 2014 REF with the outcome that would have resulted had the publications that were submitted been evaluated, when possible, with the VQR bibliometric algorithm used in the corresponding exercise for Italian universities.

We are keenly aware of the rough and approximate nature of our analysis, whose aim is chiefly to highlight a possible route to be followed in light-touch, cost-effective evaluation, which could thus be repeated more frequently, rather than a suggestion that the measures that we obtain are an accurate description of the relative standing of the UK institutions in the various subject areas. The extreme closeness of the outcome, especially when comparing size-sensitive measures, is striking. That the two mechanisms produce assessment much closer for units of assessment than for the 'representative researcher' of a department, also suggests that a bibliometric algorithm should be used with a large pinch of salt, or not at all, for individuals' assessment, and should instead be limited to a higher level of aggregation such as the department or the university. In terms of the UK current REF assessment approach, this would free the time spent by panel members in performing a task that may be performed by an algorithm just as effectively, allowing them to devote more attention and effort to the evaluation of the impact outside academia and the department research environment, as well as outputs such as books or unpublished papers, which do not lend themselves to an algorithm-based evaluation. In practical terms, of course, this would obviously not suggest changing the rules this late in the day, but rather to using the REF2021 submission for a dummy run, designed to be compared with the actual evaluation once it is completed, in order to gather detailed information on the feasibility of assisting or partially replacing, in future assessment exercises, the peer review evaluation with a less costly bibliometric mechanism.

Of course, the nature of the research output might itself be affected by the manner in which it is measured, in a coarse macroscopic version of the Heisenberg Uncertainty Principle. While the Italian VQR did not make public in advance the list of journals that would guarantee a high score, in addition to the fact that the evaluation of each output depended also on the idiosyncratic citation counts, the criteria that would determine the assessment were known, and this would inevitably inform a researcher's choice of journals providing incentives to publish mainly in these outlets, even though they might not be the most suitable ones for their research and even though this might hamper scientific innovation. This effect could be particularly strong for early career researchers, who might decide or be persuaded to submit their work to less prestigious journals, rather than submitting outputs subject to the potential uncertainty of a peer review, for example in the form of working papers.[16]

ACKNOWLEDGMENTS

The paper was presented at the Economics of Science workshop (Bordeaux), the 34th EEA conference (Manchester) and the ISSI conference (Rome). We thank the participants for comments. We thank Anne-Wil Harzing for several comments on an earlier draft. We are also grateful to editor Steve Machin and three referees of this journal for their several helpful suggestions. We thank ANVUR, the Agenzia Nazionale per la Valutazione delle Universitá e della Ricerca, for supporting this research project. The information and views set out in this paper are those of the authors and do not reflect the official opinion of the European Union or of ANVUR. Neither the European Union institutions and bodies nor ANVUR, nor any person acting on their behalf, may be held responsible for the use that may be made of the information contained herein.

NOTES

1. Research England, the Scottish Funding Council (SFC), the Higher Education Funding Council for Wales (HEFCW), and the Department for the Economy, Northern Ireland (DfE).
2. To take a specific example, the output of Unit of Assessment 18 (Economics and Econometrics) for the University of Nottingham, available at http://results.ref.ac.uk/Results/BySubmission/1564 (accessed 14 April 2021), was assessed as follows.

	% of the submission meeting the standard for				
	4*	3*	2*	1*	0*
Overall	18	71	10	0	1
Outputs	19.7	65.3	14.2	0	0.8
Environment	12.5	87.5	0	0	0
Impact	18	74	8	0	0

3. As the Economics and Econometrics panel's final report notes, a full one-quarter of the outputs that they assessed were submitted as part of an institution's submission to the Business and Management panel, and sent to them for assessment by the latter. This included outputs from 15 institutions each submitting 30 or more outputs referred to the Economics panel. See www.ref.ac.uk/2014/media/ref/content/expanel/member/Main%20Panel%20C%20overview%20report.pdf (accessed 14 April 2021).
4. The problem of strategic submission was probably less prominent in the 2014 REF than in the previous exercises, when the funding was proportional to the product of the number of full-time equivalent staff submitted and the average quality of their research. In the past exercises, submitting an additional, weak, researcher could have lowered the department average and hence the funding as well as the prestige. The change to the funding formula for the 2014 exercise described in detail in equation (3) below was intended to soften the trade-off and induce universities to submit all their research staff. Anecdotal evidence suggests, however, that the desired effect was not achieved, and the rules have changed again for the next exercise, REF2021, when all staff involved in research will have to be submitted.
5. Detailed information of how public funds are allocated to UK universities can be found at www.hesa.ac.uk/stats-finance (accessed 14 April 2021). The full set of REF rules, the identities of the reviewers and the outcomes are all available at www.ref.ac.uk (accessed 14 April 2021).
6. Although the exact details of formula (3) were determined after the publication of the results, institutions knew the principles that would underpin it.
7. The procedure is described in greater detail in Anfossi *et al.* (2016).
8. The San Francisco 'DORA' (https://sfdora.org) movement opposes the use of impact factors for this reason.
9. There are two details that are worth mentioning when discussing the values adopted in Table 4. The first is the time overlap in the two exercises: the VQR measured citations accumulated up to 2015 of articles published in the 2011–14 period; for the REF, we looked at 2015 citations of articles published in the 2008–13 period. As a consequence, the REF articles had a longer time to be cited, and this is why we disregard the slopes used by the Italian VQR in the final year. The second detail concerns the panel that assessed their work: Italian researchers chose the panel to which they submitted their paper, without knowing in advance the slopes that the panel would have adopted; in the case of the REF, given the arbitrariness of mapping the REF research areas into the VQR research areas, we have relied on the subject area of the publishing journal, which had a correspondence into the Italian panels reconstructed by Scopus.
10. As a hypothetical example, consider *Economica*. Some Italian academics submitted to the STEM VQR research areas may have publications in *Economica*. In this case, *Economica* would have been assigned by

the VQR panel to a VQR research area, and we used the slopes chosen by that VQR research area. If instead no Italian academics in the STEM VQR research areas had submitted papers published in *Economica*, then, in view of the fact that most economics journals where Italian STEM academics published were in a mathematics research area, we assign the articles published in *Economica* and submitted to the mathematics VQR research area. And similarly for all the other outputs submitted to the REF and published in journals included in the Scopus dataset.

11. This is equivalent to assuming that the institutions were able to identify correctly the assessment criteria of the potential panels, and would submit each paper to the unit of assessment giving that paper the highest evaluation. Again, we have no reason to think that papers with different areas would be systematically concentrated in certain institutions.
12. There is a tiny discrepancy between the downloadable outputs and the headline figure of outputs assessed, with 188 outputs submitted but not included in the downloadable files. This is because the evaluation agency agreed to maintain confidentiality, for commercial or for national security reasons, for some of the outputs submitted. These are clearly not journal outputs, so their absence does not affect our analysis. Similarly, while the website landing page reports 1991 UoAs, the outputs for only 1845 could be downloaded, and, of these, 5 have submitted no research product classified as 'journal articles', and 12 have submitted only journal articles that could not be matched to the Scopus database, leaving us with 1828 UoAs.
13. The outputs are distributed evenly in the six years covered by the REF, with the exception of 230 outputs that have 2007 as publication date.
14. Such specialists include agricultural colleges like the Royal Agricultural University, smaller conservatoires or art schools such as the Royal Conservatoire of Scotland, Rose Bruford College, Falmouth University or Norwich University of the Arts, and smaller institutions specializing in teaching such as Southampton Solent University or the University of Bolton.
15. It should, of course, be mentioned that panel members left the room when their own institution was being assessed.
16. This happened frequently for some subject areas in the 2014 REF, the economics and econometrics unit of assessment among them, whose assessed institutions submitted 2386 journal articles and 168 working papers, some of which were assessed as 4*. See www.ref.ac.uk/2014/media/ref/content/expanel/member/Main%20Panel%20C%20overview%20report.pdf (accessed 14 April 2021).

REFERENCES

Anfossi, A., Ciolfi, A., Costa, F., Parisi, G. and Benedetto, S. (2016). Large-scale assessment of research outputs through a weighted combination of bibliometric indicators. *Scientometrics*, **107**(2), 671–83.

Baccini, A. and De Nicolao, G. (2016). Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise. *Scientometrics*, **108**(3), 1651–71.

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), 451–66.

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C. A. and Peracchi, F. (2016). Comment to: Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise. *Scientometrics*, **108**, 349–53.

De Fraja, G., Facchini, G. and Gathergood, J. (2019). Academic salaries and public evaluation of university research: evidence from the UK Research Excellence Framework. *Economic Policy*, **34**(99), 523–83.

Farla, K. and Simmonds, P. (2015). REF accountability review: costs, benefits and burden. Technical Report, Technopolis Group.

Forster, J. (2015). Report from the RSS Working Group on Research Excellence Framework (REF) league tables. Technical Report, Royal Statistical Society, London.

Harzing, A.-W. (2018). Running the REF on a rainy Sunday afternoon: can we exchange peer review for metrics? Presented at the 23rd International Conference on Science and Technology Indicators (STI 2018), Leiden, The Netherlands.

HEFCE (2009). Report on the pilot exercise to develop bibliometric indicators for the Research Excellence Framework. Technical Report, Higher Education Funding Council for England, London.

Hirsch, J. E. (2010). An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics*, **85**, 741–54.

Mryglod, O., Kenna, R., Holovatch, Y. and Berche, B. (2015). Predicting results of the research excellence framework using departmental h-index: revisited. *Scientometrics*, **104**(3), 1013–17.

Wang, L., Vuolanto, P. and Muhonen, R. (2014). Bibliometrics in the research assessment exercise reports of Finnish universities and the relevant international perspectives. Technical Report.