Have you read this? An empirical comparison of the British REF peer review and the Italian VQR bibliometric algorithm*

Daniele Checchi† Alberto Ciolfi‡ Gianni De Fraja§
Irene Mazzotta¶ Stefano Verzillo||

January 30, 2019

Abstract

This paper determines the ranking of the publications units of assessment which where submitted to the UK research evaluation carried out in 2014, the REF, which would have been obtained if their submission had been evaluated with the bibliometric algorithm used by the Italian evaluation agency, ANVUR, for its evaluation of the research of Italian universities.

Keywords: Assessment of academic research, Incentives, University funding, University ranking, Publications, Bibliometry.

*We would like to thank ANVUR, the Agenzia nazionale per la valutazione delle Università e della Ricerca, for supporting this research project. All opinions expressed in the paper are personal and do not involve the Agency. 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. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

†ANVUR Rome, Italy, and University of Milan (Italy); email: daniele.checchi@gmail.com.

‡ANVUR Rome, Italy; email: alberto.ciolfi@anvur.it.

§University of Nottingham, School of Economics University Park, NG7 2RD, UK; email: gianni.defraja@nottingham.ac.uk. Università di Roma "Tor Vergata", DEF, Via Columbia 2, I-00133 Rome,Italy. C.E.P.R., 33 Great Sutton St, Clerkenwell, London EC1V 0DX, UK.

¶ANVUR Rome, Italy; email: Irene.Mazzotta@anvur.it.

||European Commission, JRC and University of Milan-Bicocca; email: Stefano.VERZILLO@ec.europa.eu.

1 Introduction

The week before Christmas 2014, university common rooms and PR offices up and down the country were abuzz with discussions and dissections of the freshly published results of 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. The raison d'être of the exercise is twofold. One the one hand, to ensure accountability for the taxpayer's investment in academic research and persuading the public of its benefits, on the other hand to form the basis for the selective allocation of the annual "block" budget for research to institutions. The funds allocated on the basis of the results of the REF is around one quarter of all the funds transferred from the taxpayer to higher education institutions.

Following the 2008 exercise, the funding agency run 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) and 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 £246m (Farla and Simmonds, 2015), comparable to the annual budget of a medium size university, and dividing up at £4000 per academic assessed. The next exercise, planned for 2021, will also be conducted via peer review, partly because of the 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 Wang, Vuolanto, and Muhonen (2014)'s survey. 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, Gambardella, Jappelli, Nappi, and Peracchi (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 to correspondence between the two methods (see also Baccini and De Nicolao (2016) and the reply, Bertocchi, Gambardella, Jappelli, Nappi, and Peracchi (2016)). Mryglod, Kenna, Holovatch, and Berche (2015) assess the correlation between the score and the 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, Gambardella, Jappelli, Nappi, and Peracchi (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, as we explain below, 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 (2017) has shown that ranking UK departments according to the "departmental h-index" correlates to the REF power ranking at 0.97.

In detail, we assess the papers which were submitted to the UK REF, and are included in the Scopus database, using the bibliometric criteria which ANVUR, the Italian evaluation agency, used to assess the outputs submitted for the Italian evaluation exercise which assessed outputs published from 2011 to 2014. Thus there are two important differences with the literature mentioned above. Firstly, we consider all the research areas, and, secondly, we only assess journal articles submitted to the relevant panel of 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 ANVUR algorithm; there are also several other specific differences between the two evaluations (illustrated in Table 3). 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 them 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 correlation 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 0.6. 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 policy makers is the correlation in the funding that would be attributed by the two methods, which stands at 0.995 when the departments 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. The summary result of the correlation in the institutional funding is most striking: if the output submitted had been evaluated with the bibliometric algorithm used in the Italian eValuation of the Quality of the Research (VQR), with peer review assessing the rest of the institutional submission, 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.

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 such variables appear to explain some of the difference in the scores of the two assessment methods: first the size of the submission, with larger units of assessment appearing to have been slightly penalised by the REF peer review relative to the bibliometric VQR algorithm, and the number of units in the institution as a whole, universities with many departments performing a little better with the REF than they would have done with the VQR bibliometric algorithm.

This paper is organised as follows. In Section 2, we describe the REF evaluation, and in Section 3 we present the bibliometric algorithm adopted in the Italian VQR. Section 4 describes the data used to evaluate the REF journal articles, and Section 5 reports the results. A brief conclusion ends the paper.

2 The 2014 Research Excellence Framework

The REF2014 exercise evaluated the research conducted by 52,000 academic researchers associated to 1911 units of assessment in 154 higher education UK institutions. The assessment was carried out by 36 expert panels, one in each area of research, 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 full list is in Table 5 below. 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 differ in the two exercises 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 classification of topics. We will then denote as "VQR research areas" and "REF research areas" the groups of subject areas which were assessed by the 16 VQR individual panel (known as GEV *gruppi esperti valutatori*) and the 36 REF panels. In the formal analysis we index with h the subject areas and with i the research areas.

Panels assessed 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 2008-2013; (ii) the research environment, as described in words by each institution; (iii) the impact of research on the wider society, in terms of knowledge transfer and/or public engagement, as evidenced in case-study reports, numbering one per every eight researchers.

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 "quality that is world-leading in terms of originality, significance and rigour" to the worst, 0-stars "quality that falls below the standard of nationally recognised work". 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.[1]

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

[1] 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, was assessed as follows:

	% of the submission meeting standard				
	4*	3*	2*	1*	U
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

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 as department or unit, the group of academics which an institution submitted for assessment to a specific UoA (Unit of Assessment), 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" panels, respectively. And indeed, many institutions submitted the entire department of economics to the "Business and Management Studies" panel[2]. 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 on the day-to-day life 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 payroll. These considerations suggest a loose correspondence between units of assessment and departments which moreover is unlikely to be orthogonal to the quality of the research output and casts obvious 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, Kenna, Holovatch, and Berche (2015) and

[2]As the Economics and Econometrics panel's final report notes, a full one quarter of the outputs they assessed was 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. http://www.ref.ac.uk/2014/media/ref/content/expanel/member/Main%20Panel%20C%20overview%20report.pdf

Harzing (2017).[3]

Outputs can be submitted by an institutions as long as the author is employed by that institution on the REF census date, 31st October 2013, irrespectively 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.

The environment component is a written submission describing the achievements of the academic department, together with data on research grant income and PhD completions. Finally, impact is assessed by considering written 'case studies', one for every eight academics submitted, accompanied by supporting evidence which 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 it 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, and so we limit our comparison to the output component of the REF.

Unlike its Italian counterpart, the UK funding agency does not present a single score which 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

[3]The problem of strategic submission is probably less prominent that in the previous exercises, when the funding was proportional to the product of the number of FTE staff submitted and the average quality of their research: 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 (3) 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 rules have changed again for the next exercise when all staff involved in research will have to be submitted.

Table 1:
Summary statistics and cross correlations of REF performance by component.

	GPA Score	GPA Outputs	GPA Environ.	GPA Impact	Mean	St. Dev
GPA Score	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

Note: Sample size = 1828 departments submitted to REF 2014. For explanation of the components see main text. *** denotes significance at 1% level.

scores, with the proportion in each category as weight: the GPA of department i's in institution k is calculated simply as:

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

where π_{ik}^{s} is the proportion of the activity of department i's in institution k which was assessed to be of s star quality. Table 1 reports the grade point average (GPA). 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, which again has no official status. It is simply the product of the GPA by the number of staff submitted:

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

where n_{ik} denotes the number of full–time equivalent researchers submitted by institution k to panel i. Thus GPA measure the average quality, without reference to the size of the unit of assessment, which is instead taken into account by the RP measure. There is an obvious trade-off between the two: excluding a relatively weak member of staff would definitely increase the 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 formula, FS, which is 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 which is strictly linked to specific projects, universities are free to spend this funding as they wish, with no link to projects or even disciplines.[4]

When designing the funding formula the government intended to provide incentives towards high quality research, and so it gave high weight to $4^\star$ output, specifically four times higher than the weight given to $3^\star$ output, and *no weight* to output judged less than $3^\star$.[5] With the above notation, an institution's funding in year t until the following evaluation exercise is given by

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

where Φ_t is the coefficient (in the jargon the "QR unit funding"), which varies from year to year, and depends on the overall public funding for universities, and Γ_i is a research area specific weight which takes value 1.6 for STEM subjects, UoAs 1-15, 1.3 for intermediate cost research areas such as geography, architecture, sport sciences, design, music, UoAs 16, 17, 26, 34, and 35, and 1 for all other research areas.

Table 2 shows reports the correlation between these measures, indicating that the size based ones, RP_{ik} and FS_{ikt}, are fairly close but both rather different from the GPA, 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

[4]Detailed information of how public funds are allocated to UK universities can be found at www.hesa.ac.uk/stats-finance. The full set of REF rules, the identity of the reviewers, and the outcomes are all available at www.ref.ac.uk.

[5]Although the exact details of formula (3) were determined after the publication of the results, institutions knew the principles which would underpin it.

Table 2:
Correlation between possible measures of performance

	GPA Score	Research Power	Funding Formula	Mean	St. Dev
GPA Score	1			2.82	0.433
Research Power	0.377***	1		79.62	93.11
Funding Score	0.508***	0.978***	1	38.197	50.964

Note: Sample size = 1828 departments submitted to REF 2014. For explanation of the measures see main text. *** denotes significance at 1% level.

different strategies, some preferring prestige, and thus selecting only their best performers, others pursuing the funding associated with larger submissions.

The main aim of this paper is to determine degree of similarity between the two methods of assessment, the REF peer review and the Italian bibliometric measurement. To do so, we calculate the quality scores of the output component of the research activity of the UK institutions that would have resulted if the REF assessment of the outputs 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 2011/2014 period. We stress that we do *not* attempt to perform a comparison between Italian and British institutions. For this comparison to be meaningful, the assumption should hold that British departments would have made the same submission they did for the REF 2014 if they had to be assessed according to the Italian VQR rules. Given the many differences between the set of rules used in the two assessment methods, as illustrated in Table 3, this seems unlikely.

Differences between the results of the two assessment methods could spring from two sources. One the one hand there could 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 and with low citation counts. On the

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	2011-14
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 subject area
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	pre-determined*
Funding attributed to	institutions only	both institutions and departments**
Entity assessed	department/unit	individual output

Note: Summary comparison between the VQR and the REF, see text for more details. information obtained from www.ref.ac.uk/2014 (REF) and http://www.anvur.it/attivita/vqr/vqr-2011-2014/ (VQR).

* 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.

** 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.

other hand, there might be systematic difference in the submission strategy 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 having to rely on bibliometric algorithm to select the papers and the academic 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 only be suggestive, even though, compared to some of the ex-

isting literature, it covers the whole of the research carried out in the UK.

3 The VQR bibliometric algorithm.

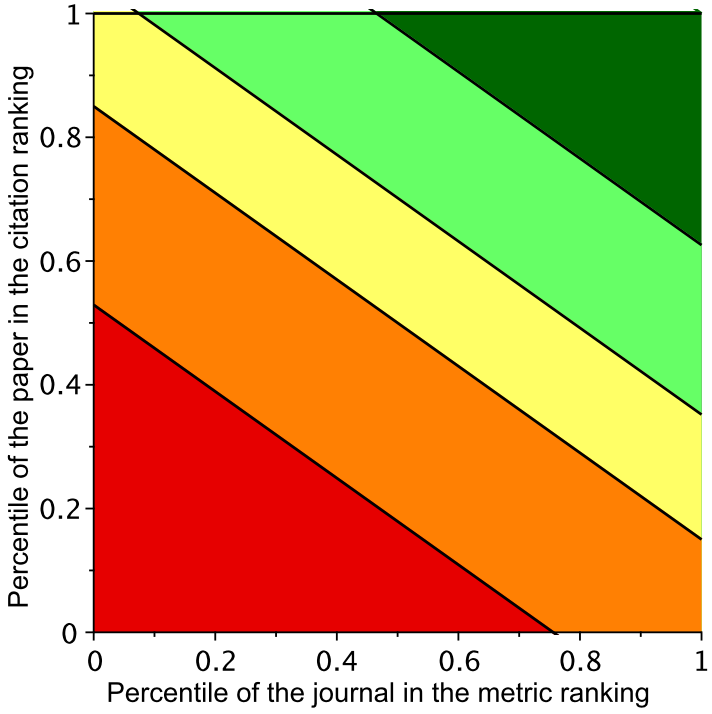
The VQR algorithm identifies a paper by four parameters: (i) the year of publication, $t = 1, \ldots, T$, (ii) the subject area, indexed by h, (iii) the number of citations at the census date, and (iv) the journal where it was published. The last two parameters are both turned into a number in $[0, 1]$ by normalising their position in an appropriate distribution, as explained in what follows. The algorithm computes the distribution of the citations obtained by all the articles published in research area h in year t; let this be denoted by $\Phi^C_{ht}(n) \in [0, 1]$. That is, $\Phi^C_{ht}(n) \in [0, 1]$ is the proportion of papers published in research area h in year t that have obtained n citations or less. Similarly for journals, where the relevant measure is the journal impact metric: $\Phi^J_{ht}(x) \in [0, 1]$ is the proportion of journals included in the Scopus database as pertaining to research area h that, in year t, had impact metric at most x.

In order to do so, it is therefore necessary to know the world distribution of citations and impact metrics at the earliest available date after the REF census date. We purchased from Scopus bibliometric information (namely the number of citations and the SCImago Journal Rank) on 1/1/2015, for each of the papers submitted to the REF; given the suggestive nature of the exercise, rather than obtaining detailed information on the world distributions of impact metrics and citations, we opted to use data made available by ANVUR, which included these distributions on 1/1/2017. This might generate a measurement error which however is systematic only to the extent that there are different trends in the citations patterns and the impact metrics of the journals where certain institutions are more inclined to publish.

In the next step of the procedure[6], the unit square $[0, 1]^2 \subseteq \mathbb{R}^2$ is divided into five subsets as shown in Figure 1 by four parallel downward sloping straight lines, in such

[6]The procedure is described in greater detail in Anfossi, Ciolfi, Costa, Parisi, and Benedetto (2016).

Figure 1:
Allocations of products to quality classes



a way that the dark green area[7] is 0.1, the light green and yellow areas are both 0.2, the orange area is 0.3, and the red area is 0.2. Simple computations determines the boundary lines; these are given by $y = a_{it} - b_{it}y$, where a_{it} is the solution in a, for $\sigma = 0.1, 0.3, 0.5, 0.8$, of:

$$1 - \max\left\{0, \frac{a-1}{b_{it}}\right\} - \int_{\min\left\{1, \frac{a}{b_{it}}\right\}}^{\max\left\{0, \frac{a-1}{b_{it}}\right\}} (a - b_{it}x)\, dx = \sigma.$$

[7]The normalisation with the percentiles ensures that the distribution is uniform in the unit square.

This solution is given by:

$$a_{it}(\sigma, b_{it}) = \begin{cases} 1 + \dfrac{b_{it}}{2} - \sigma & \text{if} \quad \sigma \leq \dfrac{b_{it}}{2} \\ 1 - \sqrt{2\sigma b_{it}} + b_{it}(1 - x) & \text{if} \quad \dfrac{b_{it}}{2} < \sigma \leq 1 - \dfrac{b_{it}}{2} \\ \sqrt{2b_{it}(1 - \sigma)} & \text{if} \quad \sigma > 1 - \dfrac{b_{it}}{2} \end{cases} \tag{4}$$

In (4) b_{it} is the slope used to assess outputs in the VQR research area i in year t: it is chosen subjectively by each panel, 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_{it} 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, and so for more recent papers the impact metric of the journal was given a higher weight. Because of these considerations, the slopes separating the areas in Figure 1 increased in absolute value with the year of publication so as to reduce the importance of citation for younger articles.

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 we have assessed. The overlap between the REF and the VQR is such that we could use the VQR slopes only for the years 2011-2013. For the other years, 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.[8]

[8]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; REF looked at 2015 citations of articles published in the 2008-13 period. As a consequence, the REF articles had longer 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 which assessed their work: Italian researchers chose the panel to which they submitted their paper, without knowing in advance the slopes which the panel would adopt; in the case of 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. For the selection

Table 4:
Slopes of trade-offs between citations and impact factor

	VQR				REF		
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 subarea				1.1	1.4	1.7
Phisics	.4	.6	.9	1.5	.4	.6	.9
Chemistry	.4	.6	.8	1.2	.4	.6	.8
Earth Sciences	.4	.6	.9	1.5	.4	.6	.9
Biology	.4	.6	.8	1.2	.4	.6	.8
Medicine	.4	.6	.8	1.2	.4	.6	.8
Agricult. and Vet. Sciences	.7	.9	1.5	2	.7	.9	1.5
Architecture	.6	.9	1.5	2	.7	.9	1.5
Civili Engineering	.7	.9	1.5	2	.7	.9	1.5
Ind. and Inform. Engineering	.4	.6	.9	1.5	.4	.6	.9
Psychology	.4	.6	1	1.5	.4	.6	1

Note: The slopes of the lines in Figure 1, for different VQR research areas and different years. The first four columns report the coefficients used in the VQR, the last three those we have used to compute the scores of papers submitted to the REF

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 which was in percentile p^C of the world distribution of citation for articles published in subject area h in year t, published in a journal whose impact metric placed it in percentile p^J of the corresponding world distribution of journals' impact

of a subject area for multi-subject journal, see below.

metrics. Then, this article's score is given by

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

where, in each row, the dependence of a_{it} on σ and b_{it} derived in (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; it is assessed as "good" (score 0.7), if it falls within 10% and 30%; it is considered "fair" (score 0.4), if it falls within 30% and 50% and as "acceptable" (score 0.1), if it falls within 50% and 80% of the world distribution. The remaining papers are labelled as "limited", and receive a score of 0.

Approximately 70% of the outputs submitted to REF are published in journals which the VQR had allocated to one or more VQR research areas. We allocated the remaining ones, for example journals in social sciences arts and humanities, to close VQR research areas, possibly more than one, by exploiting information on the frequency of publications in journals of a given Scopus subject areas by the academics submitted to a VQR research area. The entire allocation procedure was such that around 46% of the outputs submitted to the REF and contained in Scopus was published in journals which are associated to multiple VQR research areas. Depending on where they fall in the version of Figure 1 of each VQR research area, a given output could have different values of these scores. In the event, 7068 outputs, around 5% of those we assessed, were given different values by the algorithm. When this happened, we assessed the given output in all the selected VQR research areas,

and then chose the highest evaluation score.[9]

After each output was assigned to the corresponding class, the score could be aggregated by averaging or adding up all the scores for each article submitted by members of each unit assessed (department, faculty or university).[10] The corresponding score for each institution i evaluated according to the VQR algorithm is given by:

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

where π_{ik}^{s} is the proportion of the articles of institution i published in research area k to which the algorithm assigned a score $s_{VQR} = s$, $s = 1, 0.7, 0.4, 0.1$. Note of course that $\sum_s \pi_{ik}^{s} \leq 1$, but it can be strictly less than 1, as some output may score zero. In (5), we calculate the GPA with the weight vector $(4, 3, 2, 1, 0)$ used in the REF, rather that the VQR weight vector, which was $(1, 0.7, 0.4, 0.1, 0)$. The overall correlation between the measures, at 0.998, is very high.

4 The data

All the outputs submitted to the REF is available from the REF website (www.ref.ac.uk/2014) as Excel files.[11] The total number of outputs assessed is 190,962, with 81.09% of the total (154,854) journal articles, the remainder consists mainly of chapters in books (7.5%) and books (5.4%). There are many 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%).

[9] This is equivalent to assume that the institutions knew in advance 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.

[10] In fact individual researchers have access to the evaluation of their own submission.

[11] 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 accepted to maintain confidentiality, for commercial reasons or for national security on some of the outputs submitted. These are clearly not journal outputs, and so their absence does not affect our analysis.

Table 5:
Summary statistics of the paper submitted to REF 2014.

Unit of Assessment		No. inst.	output in VQR	% of REF subm.	% assessed by REF as 4*	3*	2*	1*	U
Main Panel A		121	48356	94.44	37	44	17	1	1
Clinical Medicine	1	31	13400	97.34	39	44	15	1	1
Public Health	2	32	4881	93.26	39	41	17	3	0
Allied Health Professions	3	82	10358	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	44830	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 Manuf. 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	36432	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	12202	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	64.9	27	45	26	2	0
Anthropology and Develop. 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	19.2	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
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	139468	64.20	19	45	29	5	1

Note: The columns in the Table report the name and number of the units of assessment, grouped in their respective main panels, the number of institutions submitted, the percentage of the output submitted which could be asssessed with the VQR bibliometric algorithm, and the percentage of the outputs submitted which were assessed by the REF panel as 4, 3, 2, 1 star and unclassified.

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 it was submitted, as well as standard bibliographic information such as the DOI, the publication year, the number of co-authors, the title the 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. The outputs are distributed evenly in the six years covered by the REF, with the exception of 230 outputs which have 2007 as publication date.

Scopus returned the required data for 139,847 journal articles, the remaining submissions having being published in outlets not covered by Scopus. These were books, editorials, notes, and the like. 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 chapter in books, or monographs included in Scopus, 61 were allocated in the REF published data to an anonymised UoA, and 17 had missing data which made their allocation impossible). The final tally of outputs we assessed was thus 139,468.

Table 5 presents summary statistics of the output data: as one expects, the research area with the highest proportion of outputs that can be assessed using the VQR bibliometric algorithm are those in the STEM research areas, and those, like economics, where the typical publication outlet are refereed journals.

5 Results

Our main results are reported in Table 6. The UoAs for the REF are ordered according to the percentage of output that we have been able to assess using the VQR, the fourth column in Table 5 .

Column (1) in Table 6 reports the correlation between the individual GPA scores calculated for the outputs of the various institutions which submitted to the corre-

sponding UoA using the VQR algorithm, (the formula in (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 both for the correlations between values and the rank correlations. GPA scores are averages, 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 reports the correlations in RP, and even more so in columns (5) and (6) which reports the correlations in the FS measure, the funding attributed to each unit submitted. In column (5), in the majority REF research areas this correlation exceeds 0.99 with the lowest value at 0.913, for "Music Drama and Dance". This is extremely high, considering that we could assess less than 17% of the outputs. The weighted average across REF research areas (with weights the output submitted to the REF) is 0.989. The very high values of the correlations even for REF subject areas where relatively few outputs where in Scopus journals can be explained with a correlation between the quality of the outputs submitted to journals and the quality of the books and other forms of outputs in these REF research areas: departments whose members can hit the best journals in the humanities also have members who write the best books.

The results for the rank correlation are less extreme. 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. Its lower value is likely to be due to the fact that many scores are very tightly bunched, and so small measurement errors change little in the absolute scores, but may have large impact in the ranking

The same message emerges from Figure 2. It illustrates the correlations and the rank correlations in the various units of assessment according to the various

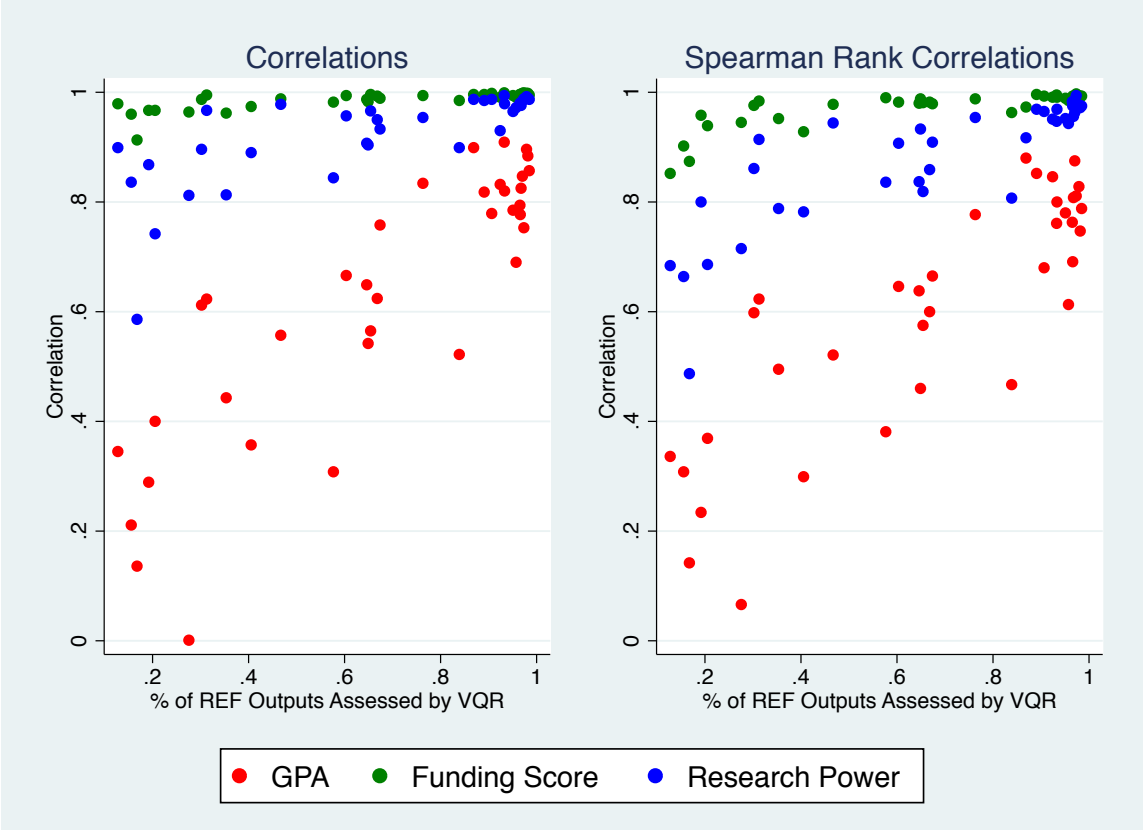
Table 6:
Correlation in the measures and the rankings

Unit of Assessment	(1) Corr GPA	(2) Spearman GPA	(3) Corr RP	(4) Spearman RP	(5) Corr FS	(6) Spearman FS
Chemistry (8)	0.857	0.788	0.987	0.975	0.995	0.993
Biology (5)	0.884	0.747	0.989	0.972	0.998	0.993
Physics (9)	0.896	0.828	0.992	0.977	0.998	0.993
Medicine (1)	0.753	0.811	0.988	0.994	0.999	0.997
Psychology (4)	0.847	0.875	0.984	0.963	0.998	0.99
Elect. Engineering (13)	0.825	0.808	0.976	0.956	0.993	0.988
Agriculture (6)	0.777	0.691	0.977	0.975	0.996	0.993
Environment (7)	0.794	0.763	0.98	0.983	0.996	0.991
Chem. Engineering (12)	0.69	0.613	0.972	0.943	0.991	0.985
General Engineering (15)	0.785	0.78	0.965	0.952	0.994	0.989
Health Professions (3)	0.82	0.8	0.979	0.969	0.996	0.991
Public Health (2)	0.909	0.761	0.994	0.947	0.999	0.995
Civil Engineering (14)	0.832	0.846	0.93	0.951	0.991	0.991
Mathematics (10)	0.779	0.68	0.987	0.965	0.998	0.993
Management (19)	0.818	0.852	0.985	0.969	0.996	0.996
Economics (18)	0.899	0.88	0.987	0.917	0.996	0.973
Sport Sciences (26)	0.522	0.467	0.899	0.807	0.985	0.963
Geography (17)	0.834	0.777	0.954	0.954	0.994	0.988
Computing (11)	0.758	0.665	0.933	0.909	0.989	0.979
Architecture (16)	0.624	0.6	0.95	0.859	0.993	0.982
Education (25)	0.565	0.575	0.966	0.819	0.996	0.981
Sociology (23)	0.542	0.46	0.904	0.933	0.983	0.988
Social Work (22)	0.649	0.638	0.907	0.837	0.987	0.98
Politics (21)	0.666	0.646	0.957	0.907	0.994	0.982
Anthr. & Development (24)	0.308	0.381	0.844	0.836	0.982	0.99
Philosophy (32)	0.557	0.521	0.978	0.944	0.988	0.978
Area Studies (27)	0.357	0.299	0.89	0.782	0.974	0.928
Media Studies (36)	0.443	0.495	0.813	0.788	0.962	0.952
History (30)	0.623	0.623	0.967	0.914	0.995	0.984
Law (20)	0.612	0.598	0.896	0.861	0.987	0.976
Modern Languages (28)	0.001	0.066	0.812	0.715	0.964	0.945
Theology (33)	0.4	0.369	0.742	0.686	0.967	0.939
English (29)	0.289	0.234	0.868	0.8	0.967	0.958
Music (35)	0.136	0.142	0.586	0.487	0.913	0.874
Arts (34)	0.211	0.308	0.836	0.664	0.96	0.902
Classics (31)	0.345	0.336	0.899	0.684	0.979	0.852

Note: Comparison between the score and the rank obtained using the VQR algorithm and the actual REF score. The horizontal line divides between UoAs where the fraction of producta assessed is above 75% and UoAs where the same fraction was below.The number in brackets after the UoA's name is the UoA's number. Pairwise correlations between each pair are respectively: 0.973***, 0.778*** and 0.903***

measures we have considered. The high correlation in institutional funding is a simple consequence of the high correlation between the scores, especially the quantity

Figure 2:
Correlations between performance scores.



Note: The diagrams report the correlations in each REF research area (LHS panel), and the correlations (RHS 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 (1), for the GPA, in (2), for the research power, and in (3), for the funding score.

based funding scores in the two methods of assessment, illustrated by the green dots.

While we stress once again the highly stylised nature of our computations, it might nevertheless be intriguing to verify, along the lines of Harzing (2017), 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 the envelope calculation finds some justification in the observation that funding is allocated to institutions, *not* departments, and so systematic errors in the funding attributable to different units in the same university may cancel out in the overall institutional funding. We do this computation only for the output

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 adding them up for all the UoAs which each institution submitted, with two different values of the "output" performance, one obtained with the VQR assessment one with the peer review assessment. (3) can be written as

$$\sum_{i \in I_k} n_{ik} \Gamma_i \sum_{s=3,4} 4^{s-3} \left(0.65 \pi_{ik}^{s,OUT} + 0.15 \pi_{ik}^{s,ENV} + 0.2 \pi_{ik}^{s,IMP} \right),$$

where $\pi_{ik}^{s,X}$ is the proportion of activity X submitted by unit i in institution k assessed to be of quality s-star, $s = 3, 4$, with X taking values OUT, output, ENV, environment, and IMP, impact, Γ_i the cost adjustment parameter taking values 1.6, 1.3 or 1, as explained above, and I_k is the set of units of assessment submitted by institution k. The correlation between the levels of funding with the two methods is 0.9997, both when all units of assessment are taken into account and when only those where at least 75% of the outputs could be assessed with the VQR algorithm. Obviously some of the correlation is due to the fact that the environment and impact components are the same in the two terms, but if these are removed, the correlation between an institution's portion of the funding due to the output component and the same portion when outputs are assessed with the VQR algorithm is still extremely high, at 0.9940, or 0.9937 when considering only the units of assessment where at least 75% of the outputs could be assessed with the VQR algorithm.

We end the paper by trying to uncover whether there are any links between any discrepancy in the two measures, our calculation using the VQR bibliometric algorithm and the REF peer review evaluation, and observable characteristics of institutions and departments. We are well aware that there is no possibility to establish any causal effect, and so the result presented in Table 7 which reports the estimated

coefficients for various specifications the following equation. (6) below.

$$\Delta_{ik} = \beta_0 + \beta_1 n_{ik} + \beta_2 N_k^U + \beta_3 N_{ik}^M + \beta_4 p_{ik} + \beta_5 w_k + \phi_i + \psi_t + \epsilon_{ik}, \tag{6}$$

where Δ_{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: $\Delta_{ik} > 0$ 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 reviewer.

In the upper part of Table 7 we include all the REF research areas. In the lower part, we restrict the sample to the REF research areas where the percentage of outputs which we were able to assess exceeded 75%.

On the right-hand side of (6), we include, n_{ik}, the number of academics submitted: this might affect the submission with the idea that a larger department might have more resources to devote to preparing the submission (for example, as some departments did, might hire an external reviewer to assist them). N_k^U and N_{ik}^M are the number 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 submission it might be easier for an institution to submit academics tactically to different panels with the aim to improve their return. A further variable we include which varies at institution level is w_k, the salary of the head of the institution (usually called Vice-Chancellor), for the year preceding the REF.

We also include a dummy p_{ik} indicating that institution k had one of its academics as a panel member for the REF research area i. This might be a variable associated with systematic difference 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 citation or

appearing a less prestigious journal, but with some characteristics which made more likely to be evaluated highly by the panel.[12] Finally, we include REF research area fixed effect, ϕ_i, and four dummies to characterise the "university type" ψ_τ: De Fraja, Facchini, and Gathergood (2016) divide all UK institutions in different types (i.e. "Russell", "1994 group" etc.): they suggest that they might have different experience and different attitudes to research. Table 7 suggests that there is very little explanatory power from any of the variables, and in the cases when we do, 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 score and ranking between departments seems to be due mostly to casual factors.

6 Concluding remarks

We have performed in this paper 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 which were included submissions been evaluated, when possible, using the VQR bibliometric algorithm used in the corresponding exercise for Italian universities.

While 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 rather than a suggestion that the measures we obtain are an accurate description of the relative standing of the UK institutions in the various subject areas, we find the closeness of the outcome, especially when comparing size sensitive measures, strongly suggestive that the method could be used to assess the publications at least for the research areas where the main outlet are refereed journals.

[12] It should of course be mentioned that panel members left the room when their own institution was being assessed.

Table 7:
Determinants of the difference in scores between VQR and REF.

	GPA		Research Power		Funding Score	
Dependant Variable:	Full	Restricted	Full	Restricted	Full	Restricted
Δ (VQR-REF)	Sample	Sample	Sample	Sample	Sample	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	1,732	803	1,676	801	1,731	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	1,732	803	1,732	803	1,732	803
R-squared	0.064	0.015	0.083	0.045	0.188	0.151

Note: 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, (1), the log of the research power, (2), and the log of the funding score, (3). The restricted sample include only the UoAs where the VQR algorithm could assess at least 75% of the outputs submitted (those above the line in Table 6). The lower part of the table repeats the OLS regression using the rank instead of the score or its log.

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. A statement that only journal articles will be considered worthwhile output for assessment would obviously direct academics to try to publish mainly in these outlets, even though they might not be the most suitable ones for their research. This

effect could be particularly strong for early career researchers, many of whose outputs were submitted in the form of working papers in some subject areas (in the economics and econometrics unit of assessment, institutions submitted 2386 journal articles and 168 working papers[13]), and who might decide or be persuaded to submit their work to less prestigious journals, rather than risk being unable to submit outputs which the rules deem of lower quality.

References

ANFOSSI, A., A. CIOLFI, F. COSTA, G. PARISI, AND S. BENEDETTO (2016): "Large-scale assessment of research outputs through a weighted combination of bibliometric indicators," *Scientometrics*, 107(2), 671–683.

BACCINI, A., AND G. DE NICOLAO (2016): "Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise," *Scientometrics*, 108(3), 1651–1671.

BERTOCCHI, G., A. GAMBARDELLA, T. JAPPELLI, C. A. NAPPI, AND F. PERACCHI (2015): "Bibliometric evaluation vs. informed peer review: Evidence from Italy," *Research Policy*, 44(2), 451–466.

——— (2016): "Comment to: Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise," *Scientometrics*, 108, 349–353.

DE FRAJA, G., G. FACCHINI, AND J. GATHERGOOD (2016): "How Much Is That Star in the Window? Professorial Salaries and Research Performance in UK Universities," Discussion Paper 11638, CEPR Discussion Paper.

FARLA, K., AND P. SIMMONDS (2015): "REF Accountability Review: Costs, benefits and burden," Discussion paper, Technopolis Group.

[13]Some of which were assessed as 4*: http://www.ref.ac.uk/2014/media/ref/content/expanel/member/Main%20Panel%20C%20overview%20report.pdf

FORSTER, J. (2015): "Report from the RSS Working Group on Research Excellence Framework (REF) League Tables," Discussion paper, Royal Statistical Society, London, UK.

HARZING, A.-W. (2017): "Running the REF on a rainy Sunday afternoon: Do metrics match peer review?," www.harzing.com.

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

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, 741–754.

MRYGLOD, O., R. KENNA, Y. HOLOVATCH, AND B. BERCHE (2015): "Predicting results of the research excellence framework using departmental h-index: revisited," *Scientometrics*, 104(3), 1013–1017.

WANG, L., P. VUOLANTO, AND R. MUHONEN (2014): "Bibliometrics in the research assessment exercise reports of Finnish universities and the relevant international perspectives," Discussion paper.