Journal : SmallCondensed 11192	Article No : 3517	Pages : 17	MS Code : 3517	Dispatch : 25-5-2020

Scientometrics
https://doi.org/10.1007/s11192-020-03517-2

Convergence or polarisation? The impact of research assessment exercises in the Italian case

Daniele Checchi[1] · Irene Mazzotta[1] · Sandro Momigliano[2] · Francesco Olivanti[3]

Received: 11 January 2020

Abstract

Two research assessments with an impact on university funding have taken place in Italy, covering the periods 2004–2010 and 2011–2014. After correcting grading schemes in order to grant comparability across the two exercises, we show that university final scores exhibit some convergence. We find that convergence is largely due to changes in the relative productivity of researchers who participated to both exercises as well as to hiring and promotions occurred between the two exercises. Results are confirmed even when we equalise the number of products across the two exercises. When we consider departments within universities, we still find convergence, though the structure and composition of departments is not strictly comparable, because mapping researchers involves some arbitrariness. These results suggest that convergence reflect genuine changes in the behaviour of researchers and in the strategies of assessed institutions, induced by incentives created by the national research assessment exercises.

Keywords Research assessment · Productivity · Convergence

A1 The opinions expressed in the paper are personal and do not involve the institutions of affiliation.

A2 ✉ Daniele Checchi
A3 daniele.checchi@unimi.it

A4 Irene Mazzotta
A5 irene.mazzotta@anvur.it

A6 Sandro Momigliano
A7 sandro.momigliano@bancaditalia.it

A8 Francesco Olivanti
A9 francesco.olivanti@osservatori.net

A10 [1] Italian National Agency for the Evaluation of Universities and Research Institutes (ANVUR), Via
A11 Ippolito Nievo 35, 00153 Rome, Italy

A12 [2] Economics and Statistics Department, Bank of Italy, Via Nazionale 91, 00184 Rome, Italy

A13 [3] Digital Innovation Observatories, School of Management of Politecnico di Milano, Via
A14 Lambruschini 4b, 20156 Milan, Italy

Journal : SmallCondensed 11192 | Article No : 3517 | Pages : 17 | MS Code : 3517 | Dispatch : 25-5-2020

Introduction

Centrally organized research evaluations have been adopted in several countries, both to measure research quality in higher education institutions and as a basis for the allocation of funding across institutions. Much attention has been given to evaluating whether such schemes have increased the quality and quantity of research. Working on a sample of 31 countries over the period 1996–2016, Checchi et al. (2019b) have shown that performance-based funding systems (PBFS) increase the number of publications after their introduction, though this effect is only temporary and fades after a few years. Looking at the scientific impact, PBFS display a negligible effect on excellence as measured by the share of articles published in top journals, irrespective of the type of assessment adopted. On the contrary, PBFS have some influence on average research quality, as measured by the number of citations per paper normalised with respect to the field.

Italy is among the countries adopting a PBFS at the entry of the present century. After a trial exercise, where participation was voluntary and there were no financial implications,[1] in 2011 a first nation-wide research assessment (VQR1-*Valutazione della Qualità della Ricerca*) was launched, covering the research activity published in 2004–2010. A second assessment (VQR2) followed 5 years later, covering 2011–2014. Both evaluations were organized under the responsibility of an independent agency (ANVUR; *Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca*) established in May 2011.[2] The exercises adopted the same approach, combining peer review and bibliometric methods. The evaluation of the research products was carried out by experts panels, grouped according to research field expertise (14 fields in the first exercise, 16 in the second). Their number was 450 during the first exercise and 436 during the second, with an overlap of 61 experts who participated to both exercises. In each exercise, the expert panels relied on approximately 15,000 external reviewers.[3] Both exercises were highly publicized, making a noteworthy impact on the reputation of institutions involved. They also had a direct impact on a significant fraction of the public funding of universities: almost one fourth of public funding to public universities (approximately 1.5 billion euro) is distributed according to the evaluation outcome.[4]

[1] The first trial exercise (VTR-*Valutazione Triennale della Ricerca*) was organized by an ad-hoc committee (CIVR-*Comitato di Indirizzo per la Valutazione della Ricerca*) and covered the period 2001–2003. Universities and research centres could submit up to half of their research staff and were free to choose the number of research products to be assessed. This ended up with many universities proposing papers by their best researchers only, while others adopted alternative strategies of involving all the researchers. All products (17,329, less than one fourth of the number of products evaluated in the two following exercises) were peer-reviewed (Cuccurullo 2006).

[2] A third assessment exercise (VQR3) has been called for in 2019 to cover research activity published in 2015–2019. While evaluation results are expected in 2021 (or 2022), the methodology has been significantly modified with respect to the previous two experiences: all products will be peer-reviewed; the number of products becomes variable across researchers allowing some researchers to compensate for the absence of others; products are to be weighed by the number of coauthors; the final result will be the allocation of product in merit categories whose boundaries are not predefined. This makes these future scores not commensurable with the scores obtained during VQR1 and VQR2 that are studied in the present paper.

[3] For an overview of the first exercise and of its results, see Ancaiani et al. (2015). The final reports of the first and the second exercises were published in 2013 and 2017 and can be downloaded from the ANVUR website (www.anvur.it).

[4] The Italian research assessment exercises have evaluated universities and public research entities, each group competing for the allocation of different sources of funds. Since research entities are more heterogeneous (they are specialised in different research fields and are unevenly distributed across the nation), we focus on the assessment of universities only.

Journal : SmallCondensed 11192 | Article No : 3517 | Pages : 17 | MS Code : 3517 | Dispatch : 25-5-2020

Over the years, various papers have criticized the Italian VQRs. Most of the critics focused on the first exercise (VQR1), but their arguments could easily been extended to the second one (VQR2) given the similarities between the twos.

Baccini (2016) and Baccini and De Nicolao (2016) criticized the "evaluative mix", i.e. having research areas (mostly STEM) evaluated through bibliometric indicators while research areas (mostly SSH) assessed through informed peer review. Given this heterogeneity, the possibility of obtaining high scores would be unevenly distributed across fields, rendering the final outcome hard to interpret. Moreover, since the joint distribution of citations and impact factors varies among bibliometric research fields, this would also introduce lack of comparability between and within research areas. Given the different "disciplinary mix" characterizing universities and departments, it would also be impossible to compare results across institutions and departments. Different conclusions were contained in Cicero et al. (2013) and Bertocchi et al. (2015), both claiming that there is a fundamental agreement between the results obtained by bibliometric indicators and peer review.

Some critics discussed the design of the bibliometric algorithm adopted in VQRs for STEM research, illustrated in Ancaiani et al. (2015). It is based on locating an article in the joint world-level distribution of citations and journal impact factors for each research field. Abramo and D'Angelo (2015) question the use of the journal impact factor, which would be a better predictor than citations only for very short citation windows (less than 2 years). More generally they criticize the use of percentile standing within the global distribution as the evaluation benchmark—instead of rescaling each publication's citations in terms of a domestic reference distribution—, which penalizes groups involved in catch-up research or in fields where the nation in question has strategic interests. Using an alternative indicator, the Fractional Article Impact Index (which also corrects for the number of coauthors), the authors show that roughly half of the top universities under VQR criteria would have not been at the top of the rankings on the basis of their global productivity, and the general ranking would have changed significantly. On the contrary, Checchi et al. (2019a) have used the VQR algorithm to evaluate the papers submitted in 2014 to the British REF, finding a rank correlation greater than 0.80 with the country ranking based on GPAs obtained from peer review.

A more general criticism (which extends to many evaluation systems, including the British REF) is that VQRs do not evaluate the entire research production of each author within the period, but only a limited subset of it due to time and money constraints imposed by peer review. Abramo et al. (2014) argue that this choice does not allow computing full productivity, jeopardizes the robustness of the peer review and poses the risk of inefficient selection of products submitted by individual researchers.[5] As a result, they suggest extending the bibliometric evaluation to all research products.

Lastly, some criticisms dealt with institutional aspects of the VQR process, among which the excessive discretion of the expert panels; the lack of full transparency of the evaluation, since the datasets were not made public; the partial information received by

[5] Using three institutions as case studies, the authors focus on the third aspect, arguing that there is a high degree of heterogeneity among institutions and researchers in the ability to select the "best" products, with a potential impact on the rankings. For STEM (the only field where the automatic evaluation of product can be applied), the results indicate a worsening by 23–32% of the maximum score achievable, compared to the score from an efficient selection. About the inability of fully understanding the complexity of the scoring system based on the VQR algorithm see also Baccini (2016).

universities on the rankings of their research staff, which limits its use for internal selective funding (Abramo et al. 2014; Baccini 2016).

The present paper is not intended to answer to previous criticisms, but is focused on a narrower demand: does the relative performance of the players (the universities) change after the experience of a first assessment? We exploit the strict similarities between VQR1 and VQR2 to investigate whether universities (or departments) who ended up at the bottom/top of the initial distribution were able to change their score in the national distribution 5 year later. Our exercise is very similar to Buckle et al. (2020) who examines whether the introduction of the New Zealand PBRF produced convergence or divergence in measured research quality across universities and disciplines between the 2003 and 2012 assessments. As in their paper, we initially inspect whether the dispersion of scores among universities/department declines between two exercises, finding a significant reduction. We then fit a standard model of mean regression (the so-called β-convergence), which is not rejected by the data.

This result is not neutral in the debate over merits and limits of PBRF. Leaving aside the issues of what should be the appropriate indicator of performance (whether including/excluding other universities' goals, like teaching or knowledge transfer) and whether one should/should not take into account differences in resource endowment, *finding evidence of convergence* suggests that *the scheme may have possibly elicited better recruitment decisions and additional effort from researchers in universities at the bottom of the distribution*.

The policy implications of our findings are particularly relevant in the Italian context, where the performance ranking of universities shows a clear geographical pattern, with Northern universities performing better than Central Italy universities, which in turn overcome Southern ones (Viesti 2016). They are also in contrast with the claim that a performance-based funding system, given the large dispersion in research quality within and between institutions in different regions, is likely to foster further divergence and inequality in the Italian higher education system (Abramo et al. 2016; Grisorio and Prota 2020). On the contrary, our conclusions suggest that the performance-based scheme does not necessarily harm the system and may have possibly given a positive contribution to it.

The paper is organised as follows. The next section introduces the data and discusses the harmonisation strategy between the two exercises. "A test for reduction in dispersion" section provides descriptive evidence of reduced dispersion of scores across the two exercises, including sample disaggregation, whereas "A test for convergence in universities' scores" section tests the convergence hypothesis. "Robustness checks" section provides robustness checks, including departmental disaggregation of the data, and "Conclusions" section concludes.

Data description and harmonization of the two exercises

The universities participating to both assessment exercises are 91 and vary significantly in size, as shown by Table 1. The largest ones count an average of 1500 researchers, against the smallest one with less than 50 academics. Overall, the number of the researchers involved exceeded 50,000. However, when comparing the two exercises, one can notice that there have been minor changes in the relevant populations, especially when considering the average size within each quartile in the middle of the distribution.

Journal : SmallCondensed 11192 | Article No : 3517 | Pages : 17 | MS Code : 3517 | Dispatch : 25-5-2020

Table 1 Researchers involved in the evaluation exercises, by university size

Quartiles	VQR 2004–2010		VQR 2011–2014	
	# universities	# researchers	# universities	# researchers
1st quartile	23	823	23	1095
2nd quartile	23	4645	23	4753
3rd quartile	23	14,075	23	13,540
4th quartile	22	35,415	22	33,188
Total	91	54,958	91	52,576

The principle of the VQRs score (called *IRAs*) is the comparison between research impact and personnel weight. This is directly inspired by the funding aim of the exercise, which calls for relative and not absolute measures. More specifically, if the average score obtained by researchers in a specific university is equal to the national average, the university share in funding would correspond to a per-capita allocation. Universities obtaining an average score above (below) the national average would receive more (less) funds compared to a per-capita distribution. In symbols, let's define v_{ji} as the score obtained by researcher i in institution j; then, the share of funds z_j going to institution j is determined according to

$$z_j = \frac{\sum_{i=1}^{n_j} v_{ji}}{\sum_{j=1}^{k}\sum_{i=1}^{n_j} v_{ji}}, \quad \sum_{j=1}^{k} z_j = 1 \tag{1}$$

where k is the number of institutions participating to the competitive allocation of funds. This index combines qualitative and quantitative dimensions, as it can easily be seen by the following transformation

$$z_j = \frac{\frac{1}{n_j}\cdot\sum_{i=1}^{n_j} v_{ji}}{\frac{1}{n}\cdot\sum_{j=1}^{k}\sum_{i=1}^{n_j} v_{ji}} \cdot \frac{n_j}{n} = \frac{\bar{v}_j}{\bar{v}} \cdot \frac{n_j}{n} \tag{2}$$

where n_j indicates the size of research staff in institution j while n indicates the national one. For an average performing institution (where the institution average mark $\bar{v}_j$ is equal to the national average mark $\bar{v}$) the fund share corresponds to its share of the research staff at the national level n_j/n (quantitative dimension); given a staff share, the higher is the research performance, the larger will be the funds received.[6]

On the other side, the use of relative measures simplifies the comparison between different research assessment exercises. However the comparison is not an easy task, as in principle it represents a counterfactual exercise. A fully homogeneous comparison would have required the evaluators assessing the products of both exercises at the same time, which is impossible. A second best alternative would have been having the evaluators of the second exercise rating also the products submitted during the first exercise: while it is in principle feasible, it would have required a significant investment in resources

[6] In practice, the algorithm used by the Italian research assessment is more complicated because of the existence of additional indicators based on PhDs programs and public engagement.

Journal : SmallCondensed 11192	Article No : 3517	Pages : 17	MS Code : 3517	Dispatch : 25-5-2020

Table 2 Grading schemes for the research assessment exercises and harmonisation adopted in the comparison of the present paper

Fraud	(self) plagiarism: -2	not assessable: 0	limited: 0
Wrong deliverables	not assessable: -1	not assessable: 0	
Absence of deliverables	missing: -0.5	not assessable: 0	
Decile 1 (in the world distribution of quality)	limited: 0	limited: 0	
Decile 2 (in the world distribution of quality)			
Decile 3 (in the world distribution of quality)		acceptable: 0.1	
Decile 4 (in the world distribution of quality)			
Decile 5 (in the world distribution of quality)			
Decile 6 (in the world distribution of quality)	acceptable: 0.5	fair: 0.4	fair: 0.4
Decile 7 (in the world distribution of quality)	good: 0.8		
Decile 8 (in the world distribution of quality)		high: 0.7	high: 0.7
Decile 9 (in the world distribution of quality)	excellent: 1		
Decile 10 (in the world distribution of quality)		excellent: 1	excellent: 1
Ex-ante mean score	*0.41*	*0.35*	*0.32*
Ex-ante standard deviation of scores	*0.43*	*0.33*	*0.36*

Gray shade indicates grades that have been modified in the harmonisation

that would have been hard to account in face of scarce resources. For this reason we are forced to assume that *evaluators involved in the two exercises adopted evaluation criteria which were on average identical*. Under such a working assumption, we are entitled to correct any differences emerging from different rules imposed to the two exercises (namely differences in the scores assignable to different rating and in the number of products to be submitted by each researcher). Given the different time length involved in each assessment, the number of products submitted in the two exercises was different: three products (journal articles, collected papers, books) in the first exercise, two products in the second. Taking into account non-deliveries, we consider 146,550 products (out of 153,749 theoretically expected) in the first VQR and 96,060 (out of 102,389 expected) in the second VQR.[7]

We start with the harmonisation of the grading schemes used in the two exercises, which were slightly different, as it can be seen from columns 1 and 2 of Table 2. The main differences are the penalisation of non-deliveries (present in VQR1 and removed in VQR2) and the more skewed distribution of potential grades at the other end of the distribution (again in VQR1—see Table 2). Our harmonisation strategy looks for an intermediate grading scale that minimises the corrections to be introduced (see column 3 in Table 2) and is based on two principles:

(i) all products below a median quality (including lack of deliverable, erroneous submission or fraud) obtain a zero score;
(ii) for the products above the median, those graded in the first exercise were randomly reassigned to keep the boundaries set in the second one.

The first correction reduced the lower tail of the first exercise and the dispersion of its scores, while the second correction produces the opposite effect in the upper tail (see final

[7] The two VQRs dealt with a larger number of products (179,280 and 114,431 respectively) because public research centers were also assessed. However, since they are subject to different incentives and unevenly distributed across the country, we exclude them from our analysis.

Journal : SmallCondensed 11192 | Article No : 3517 | Pages : 17 | MS Code : 3517 | Dispatch : 25-5-2020

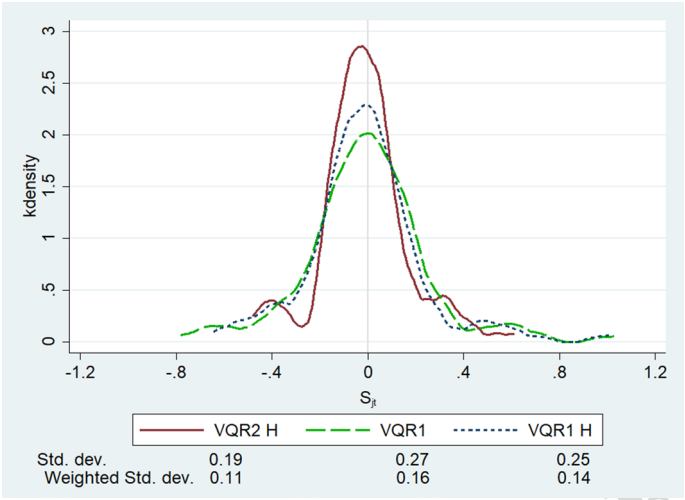

Fig. 1 Distribution of Italian universities' scores in the two research assessment exercises

rows in Table 2).[8] We have experienced with alternative distributions of the harmonised grades, without finding different results in terms of convergence (see "Robustness checks" section on robustness checks).

A test for reduction in dispersion

Once we have harmonised the grading scales of the two assessment exercises, we move to our main research question. Since we are interested in testing the convergence/divergence of universities in terms of quality of research, we compute the following statistics for each university:

$$s_{jt} = \frac{z_{jt} - \frac{n_{jt}}{n_t}}{\frac{n_{jt}}{n_t}} = \frac{\left(\frac{\frac{1}{n_{jt}} \cdot \sum_{i=1}^{n_{jt}} v_{jit}}{\frac{1}{n_t} \cdot \sum_{j=1}^{91} \sum_{i=1}^{n_{jt}} v_{jit}} \cdot \frac{n_{jt}}{n_t} \right) - \frac{n_{jt}}{n_t}}{\frac{n_{jt}}{n_t}} = \frac{\left(\frac{\bar{v}_{jt}}{\bar{v}_t} \cdot \frac{n_{jt}}{n_t} \right) - \frac{n_{jt}}{n_t}}{\frac{n_{jt}}{n_t}} = \left(\frac{\bar{v}_{jt}}{\bar{v}_t} - 1 \right); \quad j = 1, \ldots, 91; \quad t = 1, 2, \ldots \tag{3}$$

where s_{jt} can be interpreted as the deviation of university j from the mean grading in the research assessment exercise t. A positive score indicates that the university obtains an above-mean grading, while a negative value implies a below-mean performance.

[8] There is a further difference between the two exercises: while the first required submitting 3 products for each member of the faculty over a period of 7 years, the second exercise required 2 products over 4 years. It is not a priori clear whether this difference may have any implications on our analysis. See the following paragraph on robustness checks.

Journal : SmallCondensed 11192 | Article No : 3517 | Pages : 17 | MS Code : 3517 | Dispatch : 25-5-2020

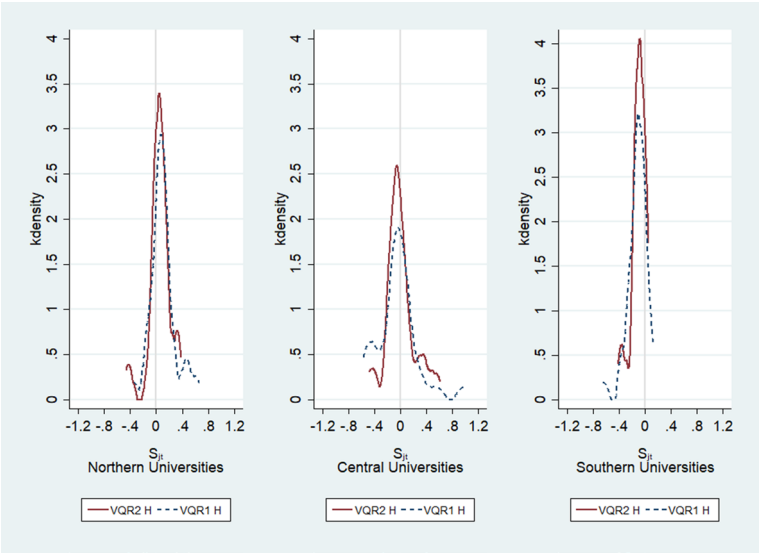

Fig. 2 Distribution of Italian universities' scores in the two research assessment exercises by geographical area

Figure 1 shows the distribution of s_{jt} in the two exercises (VQR1 2004–2010 and VQR2 2011–2014, respectively), including the harmonized scores (VQR1H and VQR2H).[9] When compared to the (non-harmonized) first exercise the distribution of the scores in the second one is clearly more concentrated around the mean. The harmonization reduces the score dispersion gap between the two exercises (their standard deviations are, respectively, 0.25 and 0.19). To verify whether the convergence that we observe is statistically significant, we apply the Fisher variance comparison test to the variances of the distribution of the harmonized exercises, obtaining a F value equal to 1.77 which is significant with p value of less than 1%.

When disaggregating the national distribution by sub-regions (see Fig. 2) one can notice interesting details. First of all the mean reversion is evident by the left-ward shift of the spike in Northern regions (left panel) as well as by the opposite shift in Southern regions (right panel). In addition, in all regions the VQR2H distribution seems more concentrated than VQR1H one.

The same analysis can be replicated over university departments. In this case let us define:

$$s_{jdt} = \frac{z_{jdt} - \frac{n_{jdt}}{n_t}}{\frac{n_{jdt}}{n_t}} = \frac{\left(\frac{\bar{v}_{jdt}}{\bar{v}_t} \cdot \frac{n_{jdt}}{n_t}\right) - \frac{n_{jdt}}{n_t}}{\frac{n_{jdt}}{n_t}} = \left(\frac{\bar{v}_{jdt}}{\bar{v}_t} - 1\right); \quad j = 1, \ldots, 91; \quad d_t = 1, \ldots, D_t; \quad t = 1, 2 \tag{4}$$

where s_{jdt} can be interpreted as the deviation of department[10] d in university j from the mean grading in the research assessment exercise t. A positive score indicates that the

[9] We omit the distribution of the non-harmonized second exercise scores because the impact of harmonization is negligible and the two curves almost coincide.

[10] Note that the total number of departments varies across universities and possibly across exercises.

Journal : SmallCondensed 11192	Article No : 3517	Pages : 17	MS Code : 3517	Dispatch : 25-5-2020

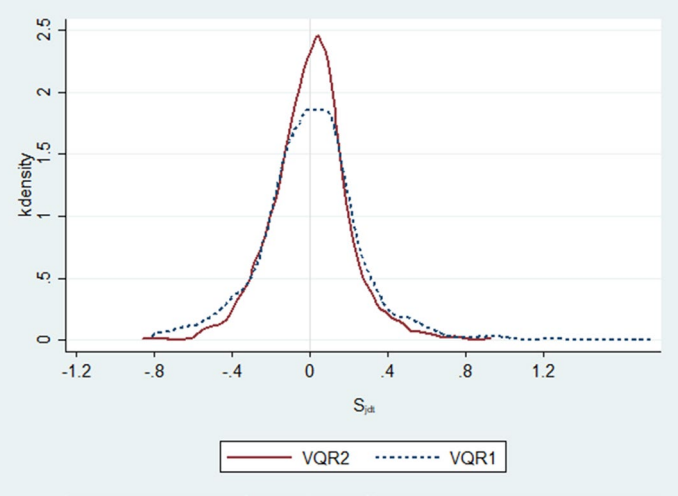

Fig. 3 Distribution of Italian universities departments' scores in the two research assessment exercises

department obtains an above-mean grading, while a negative value implies a below-mean performance. Note that, since between VQR1 and VQR2 a reform of the university system (Law 240/2010, also called "Gelmini Reform") changed the organization of the departments of all Italian institutions, the comparison of the departments in the two exercises requires mapping all researchers and departments of the pre-reform system into the new organizational structure.[11]

Figure 3 shows the distribution of s_{jdt} in the two research assessment exercises following the harmonization scheme of the evaluation scales already described for the two exercises. Once again, when compared to the first (harmonized) exercise the distribution of the scores in the second is visibly more concentrated around the mean. To verify whether the convergence that we observe is statistically significant, we perform the Fisher variance comparison test to the variances of the distribution of the harmonized exercises, obtaining a F value equal to 1.82 which is significant, with p value of less than 1%.

[11] By mapping we mean associating to each researcher, both in VQR1 and VQR2, a post-reform department (note also that researchers might have changed universities and/or department over the years). The easiest way to map old departments into new ones is to assign researchers assessed in both exercises the univocal affiliation utilized for the second VQR. However this procedure is incomplete, since a new department affiliation was still missing for 3934 researchers at the time of conclusion of VQR2 (2769 in VQR1—4.5% of the sample—and 1165 for VQR2—2.2% of the sample). This is due to delay in the completion of the reform, since some academics refused to choose a post-reform department and had to be forcefully assigned by rectors. For these cases we have proceeded as follow:

(a) in 3058 cases, we have analysed the flows of researchers within the same institution and departments from VQR1 to VQR2, and an academic has been automatically assigned to department d if more than half of her colleagues from VQR1 moved to department d in VQR2. In case of ambiguities (216 cases) we have randomly assigned these researchers to one of the possible destinations in VQR2;

(b) for 876 cases where affiliation for VQR1 was absent, we retained the researchers in the analysis of VQR2 only, and dropped them for VQR1.

Journal : SmallCondensed 11192 | Article No : 3517 | Pages : 17 | MS Code : 3517 | Dispatch : 25-5-2020

Table 3 Estimation results for linear regression model (dependent variable Δs)

Parameter	All researchers		Permanent researchers		Permanent researchers + retired/deceased ones		Permanent researchers + recruited/promoted ones	
	Value	SE	Value	SE	Value	SE	Value	SE
Intercept	−0.005	0.011	0.002	0.010	0.002	0.009	−0.004	0.012
γ	−0.383***	0.045	−0.405***	0.041	−0.372***	0.038	−0.413***	0.047

Statistical significance: ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

A test for convergence in universities' scores

Borrowing from the literature on economic growth convergence (e.g. Barro 1997), we further explore the dynamics of universities' scores with the following model:

$$\Delta s_j = \left(s_{jt} - s_{jt-1}\right) = \alpha + \gamma \cdot s_{jt-1} + \varepsilon_j \tag{5}$$

where $\hat{\gamma}$ measures the dependence from initial conditions: a negative $\hat{\gamma}$ implies convergence, i.e. regression to the mean; the closer the estimated $\hat{\gamma}$ to −1 the quicker is the convergence. A positive $\hat{\gamma}$ (or below −2) implies divergence.[12]

Table 3 and Fig. 4 shows the estimation results for the model described above in four different regressions using universities as units of analysis. In the first, we include all the researchers involved in the national research assessment (first column of Table 3; panel 4a in Fig. 4). We obtain a regression coefficient (−0.383) which is negative and highly significant, confirming the result of the F test discussed above. The result also indicates a relatively fast speed of convergence: on average, in the second exercise universities have reduced by more than a third their initial distance from the mean grading.[13]

In order to better understand the causes of the convergence we split the researchers participating to the two exercises (R_1 and R_2, respectively) into subgroups, according to the following decomposition:

$$R_{\text{perm}} + R_{1\text{only}} = R_1 \tag{6}$$

$$R_{\text{perm}} + R_{2\text{only}} = R_2 \tag{7}$$

[12] Equation (5) can be conceived as derived by the following auto-regressive process of the 1st order: $s_{jt} = \alpha + \beta s_{jt-1} + \varepsilon_{jt}$ (5′). If $0 < \beta < 1$ the process exhibits mean-reversion, i.e. it converges to a long run equilibrium given by $\bar{s} = \frac{\alpha}{(1-\beta)}$. The coefficient β captures the degree of persistence and therefore $(1-\beta)$ measures the "speed of convergence" to the long-run distribution (which in this simple framework degenerates, with all units converging to the same value). More formally, $s_{jt} = \alpha + \beta s_{jt-1} + \varepsilon_{jt}$ can be rewritten as $s_{jt} = \frac{\alpha}{1-\beta} + \beta^t s_{j0} + \varepsilon_{jt} + \beta\varepsilon_{jt-1} + \cdots$ by repeated substitution. If ε_i are iid, then $\text{Var}(s_j) = \text{Var}(\varepsilon_j)\left[1 + \beta + \beta^2 + \cdots\right] = \frac{\text{Var}(\varepsilon_j)}{[1-\beta]}$. Thus as $\beta \to 0, s_i \to \alpha$ and $\text{Var}(s_j) \to \text{Var}(\varepsilon_j)$ reaching its lowest value. On the other extreme, when $\beta \to 1$ Eq. (5′) describes a random walk, which makes it impossible to define expected moments. Given the structure of our data (the cross-sectional dimension—91—being much larger than the panel dimension—2) we cannot formally test the non-stationarity of our variable. Nevertheless, we can ensure the stationarity of our dependent variable by resorting to transformation depicted by Eq. (5).

[13] Our estimate is lower than that obtained by Buckle et al. (2020) (−0.722) with a similar strategy, but they consider a small group of universities, a selection of research fields and a longer time span.

Journal : SmallCondensed 11192 | Article No : 3517 | Pages : 17 | MS Code : 3517 | Dispatch : 25-5-2020

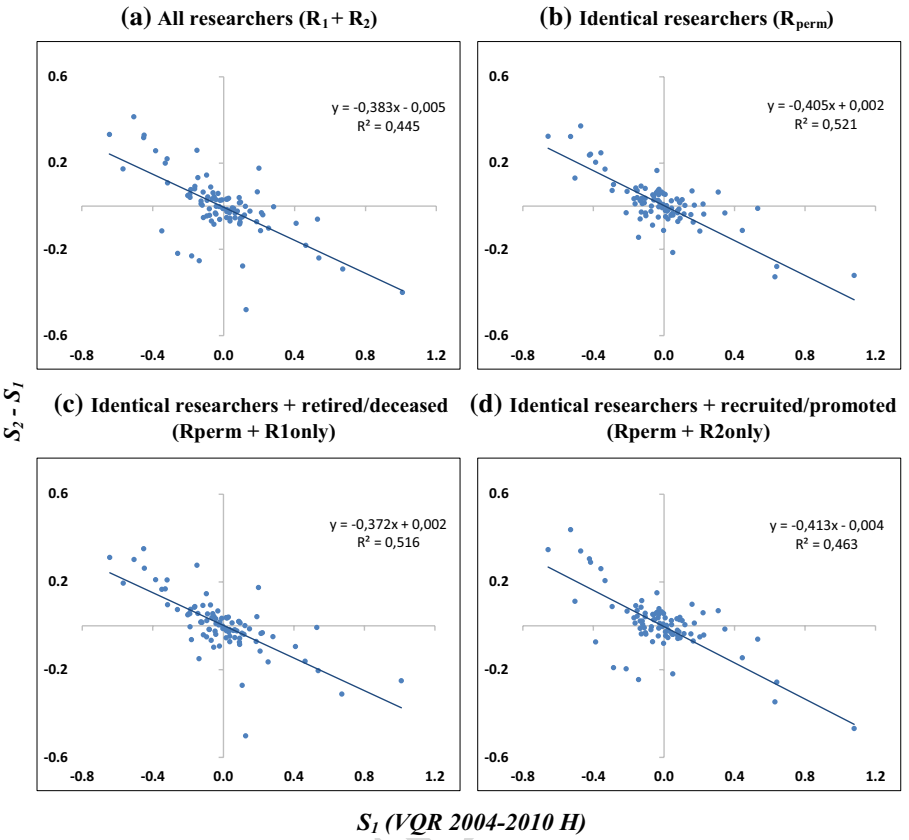

Fig. 4 Linear regression of first differences onto initial conditions. (difference between s_2 and s_1 as dependent variable and s_1 as independent variable, where s_1 = VQR 2004–2010 harmonized according to the scheme in Table 2 and s_2 = VQR 2011–2014)

where the suffix "perm" indicates the researchers who participated in both exercises with the same status (either assistant, associate or full professor), "1only" indicates those who participated only to the first exercise and "2only" those who participated only to the second or were promoted after the first exercise.

In the second regression, we restrict the analysis to the "permanent" researchers (about 45,000) who participated to both exercises and did not change status in between (R_{perm}). The regression coefficient shows an even faster convergence, indicating that the researchers reduced on average by more than 40% their gap with respect to the mean (second column of Table 3; panel 4b in Fig. 4). This rather startling result possibly stems from relative changes not only in the quality of scientific production but in a more careful selection of the research products to be submitted to the evaluation exercise. Indeed, there is some anecdotal evidence that the universities invested resources on a

Journal : SmallCondensed 11192 | Article No : 3517 | Pages : 17 | MS Code : 3517 | Dispatch : 25-5-2020

more strategic selection of products, a procedure not very frequent during the first evaluation exercise.[14]

In the third regression, we added to the "permanent" researchers those who participated only to first exercise (R_{1only}). The latter are essentially individuals (around 8200) who retired between the two exercises, as all researchers on active duty were subject to evaluation. As retirement is largely determined by age, the impact of adding this subgroup cannot be attributed to the incentives of the evaluation scheme. In this case, the regression coefficient (-0.372) is closer to zero than the two previous regressions, suggesting that the contribution of this category to convergence is either negative or very limited (third column of Table 3; panel 4c in Fig. 4).

Finally, in the fourth regression we added to the "permanent" R_{perm} researchers those who were hired or promoted (6000 individuals) after the first exercise (R_{2only}). The composition of this group essentially reflects decisions taken by the individual institutions; as the performance related scheme is largely targeted to universities, it is here that, in principle, we should see the largest impact of the reputation and monetary incentives.[15] Indeed, in this case the regression coefficient (-0.413) reaches its highest (absolute) value, suggesting a relatively large contribution of R_2only to the overall convergence (fourth column of Table 3; panel 4d in Fig. 4). These results indicate that there has been a substantial change in the behaviour of individual universities regarding hiring and promotions, with a convergence to national standards. We need to be cautious in attributing this change to the national research evaluations only, as legislation concerning hiring of academics changed substantially between the two exercises. In facts, the hiring/promotion system moved from a fully decentralized one to a centralized list of eligible candidates, among which universities could make their choices. However, since in the second period hiring and promotions were subject to the obtainment of a national qualification, the pool of candidates became a national one, thus raising the nation-wide competition among universities for attracting best candidates.[16] If therefore a low ranked university succeeded in hiring the best candidate in a field, this would have induced convergence on both sides of the distribution. On one side, it would have raised the average scientific productivity of the hiring university (since the new hired would have had a higher productivity than the incumbents); on the other hand, since it would have cream-skimmed the pool, it would have lowered the potential productivity of the best performing universities, who had no other choice than hiring second-best candidates.

[14] The result cannot be attributed to movements of researchers across institutions between the two exercises, as mobility required the opening of a position and a local competition, which were rare during the period of assessment due to the hiring freeze imposed by the central government for budgetary reasons.

[15] It should be pointed out that in Italy universities are subject to annual limits concerning the number of professors that can hire or promote.

[16] In principle any candidate was free to apply wherever she aimed to go. But local competitions were often biased in favour of local candidates, and the selecting committees were formed according to this preferred outcome. See Checchi et al. (2020).

Journal : SmallCondensed 11192 | Article No : 3517 | Pages : 17 | MS Code : 3517 | Dispatch : 25-5-2020

Table 4 Alternative harmonisation schemes

Harmonisation	Corr with H0 score (VQR1)	Corr with H0 score (VQR2)	SD (VQR1)	SD (VQR2)	*F* value
H_0 (0, .4, .7, 1)	1.000	1.000	0.25	0.19	1.77
H_1 (0, .5, .8, 1)	0.999	0.998	0.23	0.18	1.59
H_2 (−.5, .4, .7, 1)	0.782	0.991	11.51	0.41	783.21
H_3 (−.5, .5, .8, 1)	0.882	0.988	1.57	0.36	18.77
H_4 (.1, .5, .8, 1)	0.999	0.999	0.19	0.16	1.41

Robustness checks

In order to check the robustness of our results, we have investigated whether the harmonisation strategies shown in Table 2 could be responsible for the decline in variance of the universities' scores. In Table 4 the first line reproduces the results illustrated in "Data description and harmonization of the two exercises" section as benchmark H_0. With H_1 we start modifying the upper tail by using the scores adopted in VQR1 for "excellent" and "good products": given the high correlation with the previous scores, results are unaffected. In H_2 we change the bottom tail, since one of the differences between the two exercises was that VQR1 was penalising the lack of products with negative scores. Not having the corresponding information for VQR2 (which on the contrary was assigning nil scores to these cases), we are forced to attribute a negative score to all cases. In such a case the variance inflates and the differences between the two distributions widens. The H_3 case combines the previous two, by widening the two tails of the distribution, but the difference remains. Finally in H_4 the scores are rightward shifted using the grades adopted during VQR1: in such a case the difference in variance between the two distributions shrinks, but remains statistically significant. Thus the reduction in variance in VQR2 seems independent from the harmonisation scheme.

In addition to the grading scale, VQR1 and VQR2 also differed in terms of time span considered and number of expected research products. In fact, these two dimensions may have a potential impact on the variance. In particular, we expect a lower number of products submitted reducing, per se, the variability of results: if we reduce the scope of the "competition" between researchers from 3 products (VQR1) to 2 products (VQR2), the performances in the second case would converge, *ceteris paribus*, especially when considering fields using peer-review assessment. On the other hand, we expect a smaller time span, taken per se, to increase the variability of results: asking for high-quality research outputs in a shorter time window (from 7 years in VQR1 to 4 years in VQR2) makes the potential randomness of quality higher. Suppose a researcher has to select her best research outputs over a given time span: the smaller the number of products to be submitted and the longer the time span, the higher will be the probability of selecting excellent or good products. Thus, the net effect of lengthening the time span while expanding the number of product is ambiguous in terms of expected variability of product qualities.

Therefore, in addition to grading scales, we proceed by homogenizing these dimensions, in two steps. First, we select, for each researcher in VQR1, the two products with the highest score, and compare the two distributions without changing the time span considered (scenario 1). This produces a "minimum variance" scenario for VQR1. If, even in this case,

Journal : SmallCondensed 11192 | Article No : 3517 | Pages : 17 | MS Code : 3517 | Dispatch : 25-5-2020

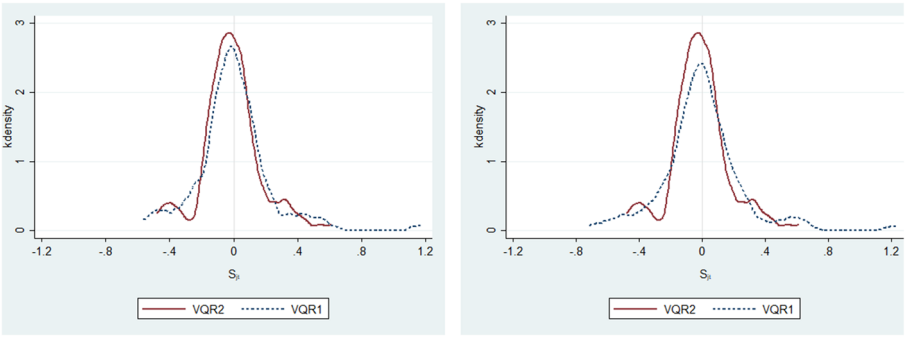

Fig. 5 Distribution of Italian universities' results in VQR1 and VQR2

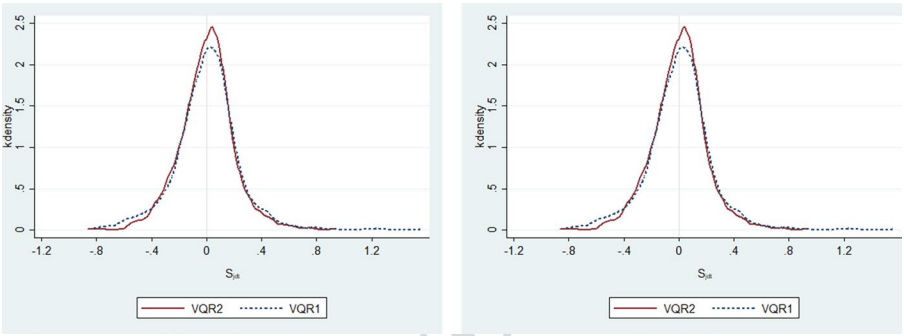

Fig. 6 Distribution of Italian university departments' results in VQR1 and VQR2

the variability of results in VQR1 is higher than that of VQR2, there is further support to the convergence result.

Second, we simulate the effect of a shorter time span for VQR1. However replicating the results of VQR1 by using the time distribution of products selected for f VQR2 (2 products over 4 years) would not produce reliable results. In fact, given the distribution of the products submitted by year of publication, reducing the time span to a 4-year window would imply having at least 10 percent of the researchers without any product to be considered (even in the most populated window, which is 2007–2010), and almost half of the population with less than two products.

Hence, we simulate the effect of a reduced window by selecting two products at random from the three products submitted. This implies picking, at least in some cases, worse products with respect to the first scenario illustrated above, hence artificially introducing some sort of divergence in the results. However, we can argue that the lower-quality products that now get picked at random are still better than any "counterfactual" second-best product that was not submitted in any 4-year window. Thus, we can interpret the variance of the distribution resulting from this simulation as a lower bound for the variance of the

Journal : SmallCondensed 11192 | Article No : 3517 | Pages : 17 | MS Code : 3517 | Dispatch : 25-5-2020

Table 5 Changes in dispersion of scores between VQR1 and VQR2

	SD (VQR1)	SD (VQR2)	Difference (%)	SD test result: $2 * Pr(F>f)$
Scenario 0	Only grading scale harmonized			
Universities	0.2393937	0.1856616	−22.4**	0.0161
Departments	0.2163716	0.1902209	−12.1**	0.0198
Scenario 1	Best two products, different time spans, and harmonized grading scale			
Universities	0.2305904	0.1856616	−19.5*	0.0533
Departments	0.1930253	0.1902209	−1.5	0.7463
Scenario 2	Two products chosen at random and harmonized grading scale			
Universities	0.2438872	0.1856616	−23.9**	0.0261
Departments	0.2764931	0.1902209	−32.4***	0.0001

Statistical significance: ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

counterfactual distribution of VQR1. The two scenarios are illustrated in Figs. 5 and 6, for universities and departments respectively.

Table 5 shows how the variability of university scores, as measured by the standard deviations of the kernel densities (Figs. 5 and 6), changes across the VQRs under the different scenarios. To verify the statistical significance of the differences across the two exercises, we perform tests on the equality of standard deviations. As already reported, there is a sizeable difference when harmonizing only for the grading scale (according to Table 2, indicated as scenario 0). When harmonizing also the number of submitted products (scenario 1, the 'minimum variance scenario'), the distribution of university results is still more concentrated in VQR2 than in VQR1: the standard deviation decreases by 19.5 percent; however, the difference between the variances is significant only at the 10% level. On the contrary, the distribution among departments vanishes. When we move to the full simulation (scenario 2); in this case, the decrease in variance between university scores in VQR1 and those in VQR2 amounts to about 24% and statistically significant with *p value* below 3%. The decline in variance among departmental scores is sizeable (32.4%) and statistically significant with *p value* below 1%. We conclude that the convergence we have described above for universities is robust also to other differences in the structure of the VQR exercise. In scenario 2, the difference in standard deviations

In order to investigate whether some research area was in general more responsive to the pressure created by the second assessment, via stricter scrutiny of products to be submitted and/or better selection of candidates to be hired/promoted, we have disaggregated the distributions by research areas.[17] Results are shown in Fig. 7. There are few cases where the decline in dispersion is evident (Medicine and Biology, and at a less extent Law), but these are counterbalanced by other cases where dispersion increases (Physics and Architecture). This indeterminacy by research field is not surprising, since universities do not contain all research fields in equal proportions, nor the research fields do represent a strategic player in

[17] Italian academics are pigeon-holed in 371 research field (*settori scientifico-disciplinari*), which are then grouped in 14 main research areas (*aree CUN*) which are used to aggregate the data shown in Fig. 7. Since VQR2 introduce the split of two areas (8 and 11), we have extended the comparison to these sub-areas.

Journal : SmallCondensed 11192 | Article No : 3517 | Pages : 17 | MS Code : 3517 | Dispatch : 25-5-2020

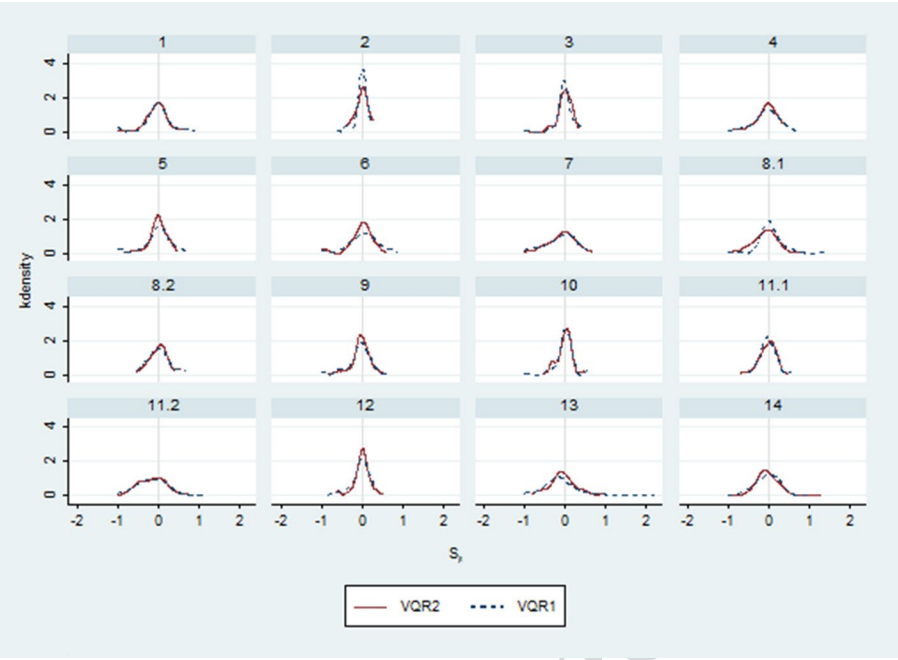

Fig. 7 Distribution of universities' scores using harmonized grading scale in VQR1 and VQR2 by research area. Note: Bibliometric sectors (VQR algorithm) = 1: Mathematics and Computer Sciences; 2: Physics; 3: Chemistry; 4: Earth Sciences; 5: Biology; 6: Medicine; 7: Agricultural and veterinary sciences; 8.2: Civil Engineering; 9: Industrial and Information Engineering; 11.2: Psychology. Non-bibliometric sector (peer-reviewed) = 8.1: Architecture; 10: Ancient History, Philology, Literature and Art History; 11.1: History, Philosophy, Pedagogy; 12: Law; 13: Economics and Statistics; 14: Political and Social Sciences

the evaluation game. This disaggregation brings support to our interpretation that universities (and departments) were the real actors in the evaluation exercise, and there is robust evidence that they changed their strategy in product selection and in hiring.

Conclusions

Performance based funding are often subject to the criticism that they produce cumulative cycles, where worse performing institutions lose money and find more and more difficult to catch up better performing ones. In this paper, we provide first evidence on this issue, comparing the results achieved by Italian universities in the two national research evaluation exercises, respectively completed in 2013 and in 2017. We find that, contrary to what expected by critics of the national evaluation exercise, the dispersion in research quality across universities has significantly fallen in the second exercise, even after correcting for differences in grading scales and in the number of products. We also find that convergence is largely due to changes in the relative productivity of researchers who participated to both exercises and to the hiring/promoting decisions of universities. The degree of convergence falls instead when we include the changes due to researchers' retirement (an event which is almost entirely determined by

Journal : SmallCondensed 11192	Article No : 3517	Pages : 17	MS Code : 3517	Dispatch : 25-5-2020

age). These results suggest that convergence may be the outcome of changes in the strategy of researchers and institutions, which may have reacted to the monetary and reputation incentives created by the national research assessments.

References

Abramo, G., D'Angelo, C. A., & Rosati, F. (2016). The North-South divide in the Italian higher education system. *Scientometrics, 109*(3), 2093–2117. https://doi.org/10.1007/s11192-016-2141-9.

Abramo, G., & D'Angelo, C. A. (2015). The VQR, Italy's second national research assessment: methodological failures and ranking distortions. *Journal of the Association for Information Science and Technology, 66*(11), 2202–2214. https://doi.org/10.1002/asi.23323.

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2014). Inefficiency in selecting products for submission to national research assessment exercises. *Scientometrics, 98*(3), 2069–2086. https://doi.org/10.1007/s11192-013-1177-3.

Ancaiani, A., Anfossi, A., Barbara, A., Benedetto, S., Blasi, B., Carletti, V., et al. (2015). Evaluating scientific research in Italy: the 2004–10 research evaluation exercise. *Research Evaluation, 24*(3), 242–255.

Baccini, A. (2016). Napoleon and the bibliometric evaluation of research: Considerations on university reform and the action of the national evaluation agency in Italy. [Napoléon et l'évaluation bibliométrique de la recherche: Considérations sur la réforme de l'université et sur l'action de l'Agence Nationale d'évaluation en Italie]. *Canadian Journal of Information and Library Science, 40*(1), 37–57.

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

Barro, R. J. (1997). *Determinants of economic growth: A cross-country empirical study*. Cambridge, MA: MIT Press.

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C. A., & Peracchi, F. (2015). Bibliometric evaluation vs. informed peer review: Evidence from Italy. *Res Policy, 44*(2), 451–466. https://doi.org/10.1016/j.respol.2014.08.004.

Buckle, R. A., Creedy, J., & Gemmell, N. (2020). Is external research assessment associated with convergence or divergence of research quality across universities and disciplines? Evidence from the PBRF process in New Zealand. *Appl Econ*. https://doi.org/10.1080/00036846.2020.1725235.

Checchi, D., Ciolfi, A., De Fraja, G., Mazzotta, I., & Verzillo, S. (2019a). Have you read this? An empirical comparison of the British REF peer review and the Italian VQR bibliometric algorithm. CEPR Discussion Paper 13521/2019

Checchi, D., De Fraja, G., & Verzillo, S. (2020). Incentives and careers in academia: Theory and empirical analysis. *The Review of Economics and Statistics.* https://www.mitpressjournals.org/doi/abs/10.1162/rest_a_00916 **(forthcoming)**.

Checchi, D., Malgarini, M., & Sarlo, S. (2019b). Do performance-based research funding systems affect research production and impact? *Higher Education Quarterly, 73,* 45–69.

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

Cuccurullo, F. (2006). La Valutazione Triennale della Ricerca - VTR del CIVR: bilancio di un'esperienza. *Analysis-Rivista di cultura e politica scientifica, 3–4,* 5–7.

Grisorio, M. J., & Prota, F. (2020). Italy's national research assessment: Some unpleasant effects. *Stud High Educ, 45*(4), 736–754. https://doi.org/10.1080/03075079.2019.1693989.

Viesti, G. (2016). (a cura di). *Università in declino.* Donzelli editore