SCIENCE, TECHNOLOGY & INNOVATION INDICATORS

Open indicators: innovation, participation and actor-based STI indicators
PARIS 2017

Research assessment in Italy: have the results of universities converged?

Daniele Checchi*, Irene Mazzotta* and Sandro Momigliano *

*daniele.checchi@anvur.it; irene.mazzotta@anvur.it; sandro.momigliano@anvur.it
Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca (ANVUR), Via Ippolito Nievo, 35 ROMA 00153 (ITALY)

ABSTRACT

Two research assessments with an impact on university funding took place in Italy, covering the periods 2004-10 and 2011-14. After adjusting the grading in order to increase comparability across the two exercises, we show that university grades exhibit a significant degree of convergence. 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 decisions of universities. The speed of convergence falls instead when we include the changes due to researchers' retirement (an event which is almost entirely determined by age). These results suggest that convergence may reflect changes in the behaviour of individuals and institutions induced by the monetary and reputation incentives created by the national research assessment.

INTRODUCTION

In Italy, the research quality of universities has been assessed in two national exercises[1], referred to the periods 2004-10 and 2011-14 respectively, under the responsibility of an independent agency (ANVUR; Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca) established in May 2011. The exercises followed the same approach, a combination of peer review and bibliometric methods.[2] 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. In particular, if the average grade of the researchers of a specific university was equal to the national average, the university share in funding would correspond to a per-capita allocation. Universities obtaining

[1] 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.

[2] The evaluation of the research products was carried out by "evaluation experts", grouped by 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 recruited about 15000 external referees. 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 found at the ANVUR website (www.anvur.it).

an average grade above (below) the national average would obtain more (less) funds compared to the per-capita distribution. In symbols, let's define v_{ji} as the mark obtained by researcher i in institution j; then, the share of funds z going to institution j has been determined according to

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

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

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

where n_j indicates the size of research staff in institution j. 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); but given a staff share, the higher is the research performance, the larger will be the funds obtained.[3]

This criterion has been criticised because it ignores other universities' goals (like teaching) and does not take into account differences in resource endowment, which can affect quality of research. Moreover, it has been suggested that it may produce cumulative vicious cycles, since worse performing institutions lose money, making it more and more difficult to catch up better performing ones.

These criticisms are particularly relevant in the Italian context, where the performance ranking of universities shows a geographical pattern, with Northern universities performing better than Central Italy universities, which in turn overcome Southern ones (e.g. see Viesti et al., . In such a framework, a crucial question, addressed in this paper, is whether the performance of these universities exhibits some sort of convergence. A positive answer would suggest that the performance-based scheme has not harmed the system and may have possibly elicited additional efforts from researchers in low performing universities and better recruitment decisions. On the contrary, a negative answer would suggest the need to modify the design of the performance-based scheme to make it more compatible with the goal of providing equal opportunity to researchers independently of their workplace.

BASIC DATA AND HARMONIZATION OF GRADES

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 50000.

The comparison of different research assessment exercises is not an easy task, as it represents a counterfactual exercise. A fully homogeneous comparison would have required the evaluators of the second exercise to assess also the products submitted during the first exercise. Given our limited goal of comparing the dispersion in the distribution of

[3] In practice, the algorithm used by the Italian research assessment is more complicated because of the existence of research areas with different evaluation standards.

universities' grades, we deemed sufficient to harmonize the grading schemes of the two exercises, which were slightly different, as it can be seen from Table 2.

Table 1 - Researchers by Italian Universities.

Quartiles	VQR 2004-2010		VQR 2011-2014	
	# universities	# researchers	# universities	# researchers
1st quartile	22	35,415	22	33,188
2nd quartile	23	14,075	23	13,540
3rd quartile	23	4,645	23	4,753
4th quartile	23	823	23	1,095
Total	*91*	*54,958*	*91*	*52,576*

In particular, we adopted the harmonisation scheme described in the last column:

i) all products below a "world" median in quality (including lack of deliverable, erroneous submission or fraud) obtained a zero score;

ii) in the current version of the paper, for the products above the "world" median, those graded in the first exercise were randomly reassigned to respect the boundaries set in the second one. We are currently working on a non-random reassignment.

The first correction reduced the lower tail of the first exercise and the dispersion of grades, while the second correction increased the dispersion (see final lines in Table 2).[4]

Table 2 – Grading schemes for the research assessment exercises and harmonisation adopted in the comparison

Research product allocation	VQR 2004-2010	VQR 2011-2014	Harmonisation
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*

TESTS FOR CONVERGENCE

Since we are interested in testing the convergence/divergence of universities in terms of quality of research, we compute the following statistics.

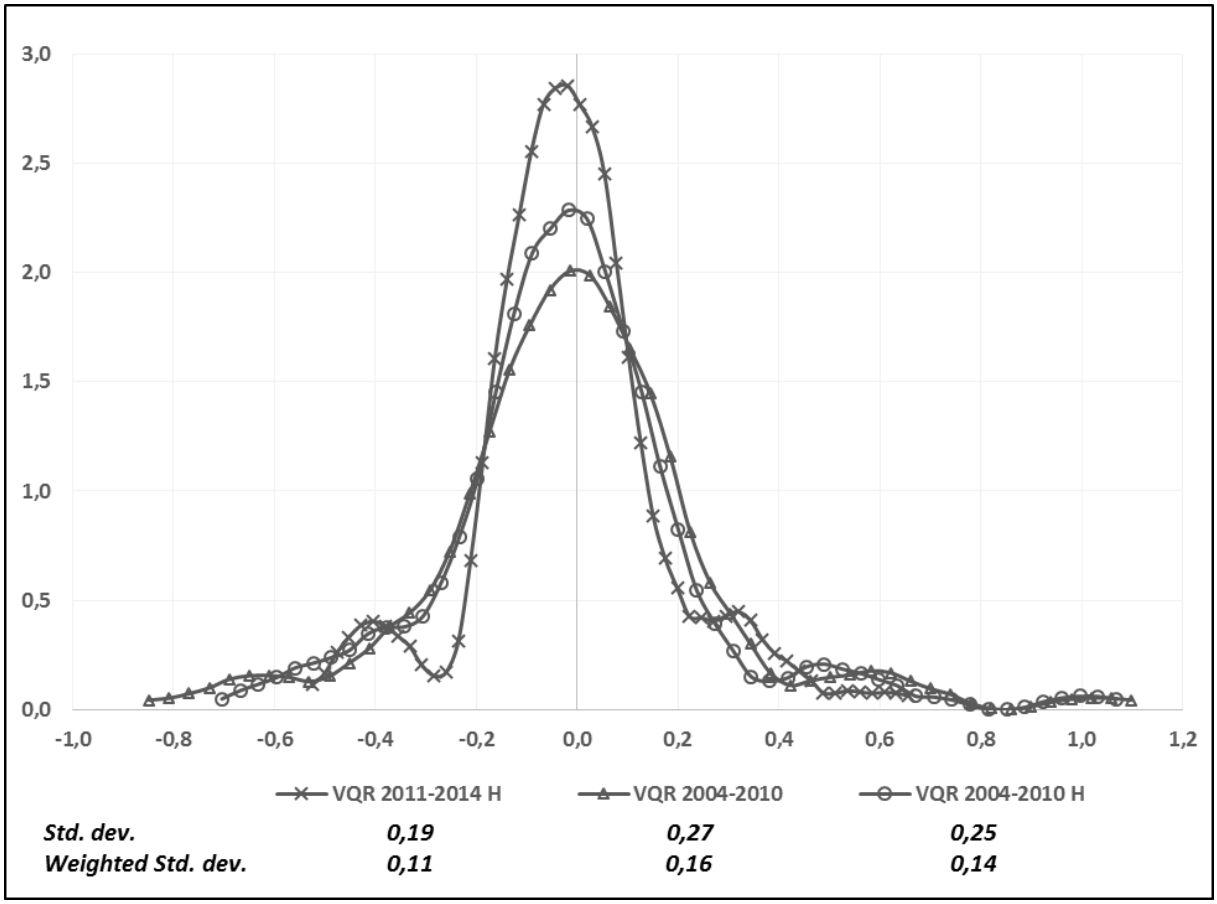
[4] There is a further difference between the two exercises: while the first required to submit 3 products per each member of the faculty over a period of 7 years, the second exercise required 2 products over 4 years. It is not a priori clear whether this difference may have any implications on our analysis.

$$s_{jt} = z_{jt} - \frac{n_j}{n} = \frac{\frac{1}{n_j}\sum_{i=1}^{n_j} v_{jit}}{\frac{1}{n}\sum_{j=1}^{92}\sum_{i=1}^{n_j} v_{jit}} \cdot \frac{n_j}{n} - \frac{n_j}{n} = \frac{\bar{v}_{jt}}{\bar{v}_t} \cdot \frac{n_j}{n} - \frac{n_j}{n} = \left(\frac{\bar{v}_{jt}}{\bar{v}_t} - 1\right) \cdot \frac{n_j}{n}; \qquad j = 1, \dots, 92; \qquad t = 1,2 \qquad (3)$$

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

Figure 1 shows the distribution of s_{jt} in the two research assessment exercises (VQR 2014-2010 and VQR 2011-14 H2, respectively) following the harmonization scheme of the evaluation scales described above. The distribution based on the non-harmonized first exercise is also shown (VQR 2004-2010); we do not report the distribution of universities' scores for the non-harmonized second exercise because in that case the impact of harmonization is negligible. Compared to the first exercise (non-harmonized) the distribution of the scores in the second is visibly more concentrated around the mean. The harmonization reduces the decline in the dispersion of the scores 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%.

Figure 1: Distribution of universities' scores in the two research assessment exercises



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

$$\Delta s_j = \left(s_{jt} - s_{jt-1}\right) = \alpha + \gamma \cdot s_{jt-1} + \epsilon_j \qquad (4)$$

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. If γ is instead positive (or it is below -2), we would be observing divergence[5].

Table 3 and Figure 2 shows the estimation results for the model described above in four different regressions. In the first, we include all the researchers involved in the national research assessment (first column of Table 3; exercise 2a in Figure 2). 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.

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.

$$R^{perm} + R^{1only} = R_1 \qquad (6)$$

$$R^{perm} + R^{2only} = R_2 \qquad (7)$$

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

Table 3. Estimation results for linear regression model (dependent variable I2-I1).

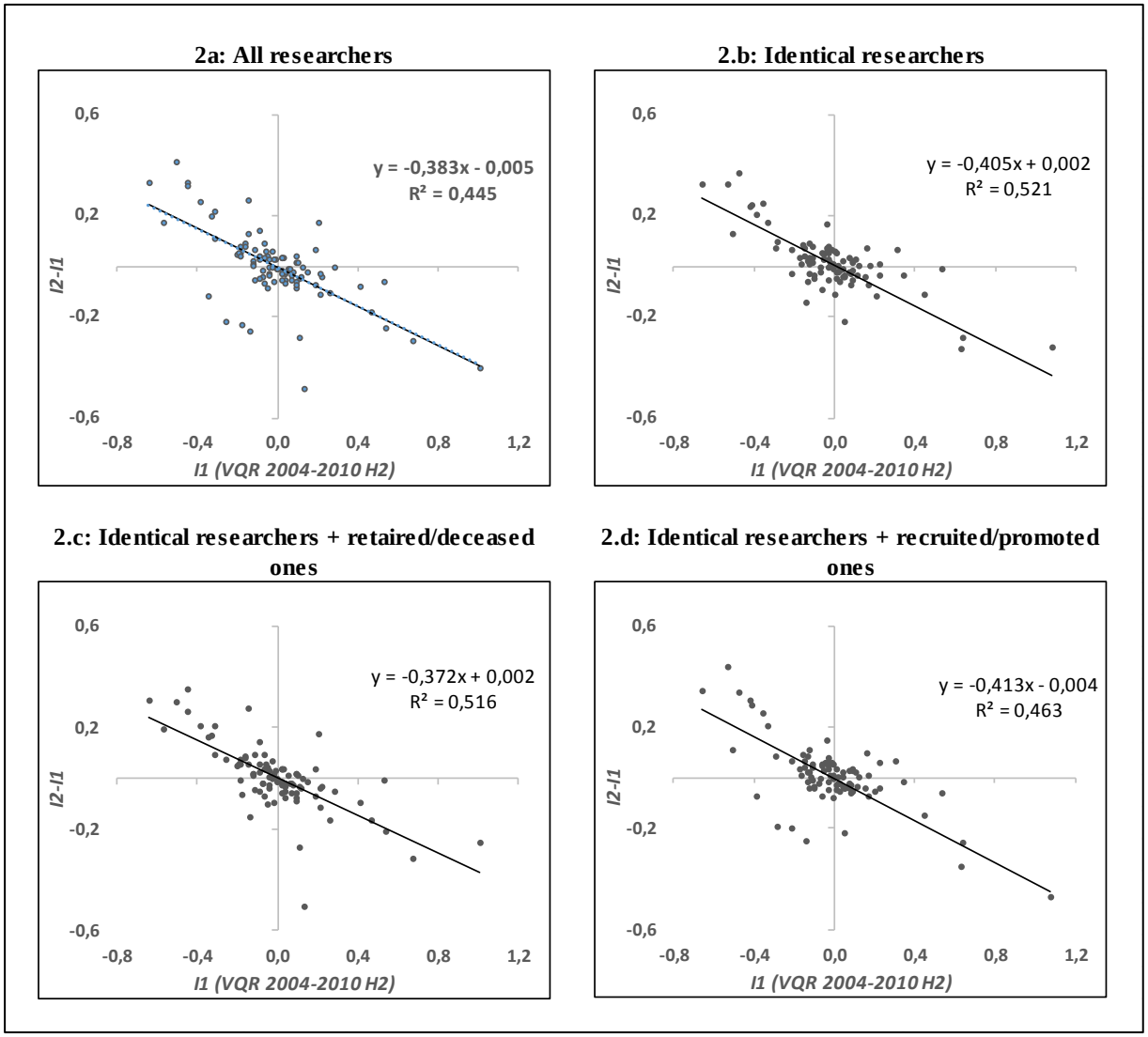
Parameter	All researchers		Permanent researchers		Permanent researchers + retired/deceased ones		Permanent researchers + recruited/promoted ones	
	Value	error	value	Error	value	Error	Value	Error
Intercept	-0.005	0.011	0.002	0.010	0.002	0.009	-0.004	0.012

[5] Equation *(4)* can be conceived as derived by the following auto-regressive process of the 1st order:: $s_{jt} = \alpha + \beta s_{jt-1} + \epsilon_{jt}$ *(5)* If $0 < \beta < 1$ the process exhibits mean-reversion, i.e. it converges to a long run equilibrium given by $\bar{s} = \alpha/(1-\beta)$. 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, $s_{jt} = \alpha + \beta s_{jt-1} + \epsilon_{jt}$ can be rewritten as $s_{jt} = \frac{\alpha}{1-\beta} + \beta^t s_{j0} + \epsilon_{jt} + \beta\epsilon_{jt-1} + \cdots$ by repeated substitution. If ϵ_i are iid, then $Var(s_j) = Var(\epsilon_j)[1 + \beta + \beta^2 + \cdots] = Var(\epsilon_j)/[1-\beta]$. Thus as $\beta \to 0, s_i \to \alpha$ and $Var(s_j) \to Var(\epsilon_j)$ reaching its lowest value. On the other extreme, when $\beta \to 1$ equation *(5)* describes a random walk, which makes it impossible to define expected moments. Given the structure of our data (the cross-sectional dimension – 92 – 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 equation *(4)*.

γ	-0.383***	0.045	-0.405***	0.041	-0.372***	0.038	-0.413***	0.047

Figure 2: Linear regression of first difference onto initial conditions

(difference between I2 and I1 as dependent variable and I1 as independent variable - I1=VQR 2004-2010 harmonized according to the scheme in Table 1; I2=VQR 2011-2014).



In the second regression, we restrict ourselves on the "permanent" researchers (around 45,000), those who participated to both exercises and did not change status in between (R^{perm}). In this case, the regression coefficient (-0,405) shows an even faster convergence to the mean, indicating that the researchers reduced on average by more than 40% their gap with respect to the mean (second column of Table 3; exercise 2b in Figure 2). This rather startling result possibly stems from relative changes not only in the quality of scientific production but also in the carefulness of the selection of the products submitted to the evaluation exercise. Indeed, there is some anecdotal evidence that the importance of carefully selecting the products was not widely appreciated in the first evaluation exercise. The result cannot be attributed, instead, to movements of researchers across institutions between the two exercises, as there has been a very small number of such movements.

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 8,200) who retired

between the two exercises, as all researchers on active duty were subject to the 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 the convergence is either nul or very limited (third column of Table 3; exercise 2c in Figure 2).

Finally, in the fourth regression we added to the "permanent" R^{perm} researchers those who were hired or promoted (6,000 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.[6] Indeed, in this case the regression coefficient (-0.413) is the highest (in absolute value), suggesting a relatively large contribution of R^{2only} to the overall convergence (fourth column of Table 3; exercise 2d in Figure 2).

In order to further investigate the role of this subgroup of researchers, we have run an additional regression, including only promoted and newly hired in both exercises. The regression coefficient is highly significant and even lower than that of the previous regression. The result (not reported in this version of the paper) suggests that there has been a substantial change in the behaviour of individual universities concerning hiring and promotions, with a fast convergence to national standards. However, we need to be cautious in attributing this change to the national research evaluations, as legislation concerning hiring of academics changed substantially between the two exercises (the system moved from a fully decentralized one to a centralized list of eligible candidates, among which universities can make their choices).

CONCLUSIONS

Performance based funding are often subject to the criticism that they produce cumulative cycles, with worse performing institutions losing money and finding it 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 two national research evaluation exercises which took place in Italy in the last decade. We find that, contrary to what is expected by critics, the dispersion in research quality across universities has significantly fallen in the second exercise, even after correcting for differences in grading scales. 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 decisions of universities. The speed of convergence falls instead when we include the changes due to researchers' retirement (an event which is almost entirely determined by age). These results suggest that convergence may reflect changes in the behaviour of individuals and institutions induced by the monetary and reputation incentives created by the national research assessment.

REFERENCES

Ancaiani, A., Anfossi, A., Barbara, A., Benedetto, S., Blasi, B., Carletti, V., Cicero, T., Ciolfi, A., Costa, F., Colizza, G. Costantini, M., di Cristina, F., Ferrara, A., Lacatena, R., Malgarini, M., Mazzotta, I., Nappi, C., Romagnosi, S. and Sileoni, S. (2015). Evaluating scientific

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

research in Italy: The 2004-10 research evaluation exercise. *Research Evaluation*, 24 (3), 242–255.

Barro, R. J. (1997). *Determinants of Economic Growth: A Cross-Country Empirical Study*. Cambridge, MA: MIT Press.