Italian Political Science, VOLUME 12, ISSUE 1, JUNE 2017

The Italian Research Assessment Exercises

Daniele Checchi
UNIVERSITY OF MILAN

The Italian experience

Italian universities have so far experienced three assessment exercises (2001-3, 2004-10 and 2011-14), which are described in details in Table 1. The fiscal law approved in December 2016 dictates that from now onwards the reference periods will be quinquennial, reducing the discretionary power so far exercised by the Ministry of Education in designing the exercise.

Table 1. The three research assessment exercises

Acronym	VTR 2001-2003	VQR 2004-2010	VQR 2011-2014
Evaluator Agency	CIVR	ANVUR	ANVUR
Output to be evaluated	Research outputs, proportional to the number of researchers – no constraints on the research areas to be covered	A fixed number of research outputs for each researcher (3 for university professors, 5 for researchers in research agencies)	A fixed number of research outputs for each researcher (2 for university professors, 3 for researchers in research agencies)
Period of reference	2001-2003	2004-2010	2011-2014
Subjects to be assessed	Compulsory for universities, voluntary for research agencies – finer assessment down to departments	Compulsory for universities and research agencies under the monitoring of Ministry of Education – voluntary for any other research agency – finer assessment down to departments	Compulsory for universities and research agencies under the monitoring of Ministry of Education – voluntary for any other research agency – finer assessment down to departments
No. of research outputs	18.000	185.000	105.000
Assessment scale	4 merit classes	4 merit classes	5 merit classes
Assessment method	peer review	bibliometric methods for hard sciences - peer review for soft sciences	bibliometric methods for hard sciences - peer review for soft sciences
Other indicators for assessment	• ownership of copyright and royalties • number of PhDs and Postdocs • degree of internationalisation (researcher mobility) • attraction of funding • ability to self-financing research projects	• attraction of funding • ability to self-financing research projects • scientific merit of newly hired/promoted personnel • degree of internationalisation (researcher mobility, co-authorship with foreign-based researchers) • improvement over the positioning during previous VTR 2001-2003	• attraction of funding • scientific merit of newly hired/promoted personnel • improvement over the positioning during previous VQR 2004-2010
External reviewers	yes	yes	yes
Panels	20 groups of evaluators	14 groups of evaluators	20 groups of evaluators
Impacts on public funding	no	yes	Yes

After an initial trial-and-error approach, the second and third exercises have been rather similar, thus consolidating a standard of evaluation, whose principles are the following:

- each assessment is intended to evaluate groups (universities, research agencies, down to departments and institutes) and not individuals (individual assessments are revealed to each researcher, but not to heads of departments, deans or chancellors);

Volume 12, Issue 1, pp. 46–54.

- the assessment considers a fixed number of products per capita/year, which should capture the best production: as such, it is closer to a monitoring exercise than to a quality assessment, revealing the excellences in a given research field;
- using current standards (1/2 product per year per university professor – currently around 52,000 – and 1 product per researcher working in a research agency – currently around 10,000) implies approximately 35,000 products per year; over a 5-year interval it sums up to 175,000 products, making some sort of automatic (bibliometric) assessment unavoidable;
- the process has been managed by groups of experts, defined according to predefined research areas (since Italian professors are pigeon-holed into 371 research fields, then grouped into 14 research areas, known as Aree CUN). Each group was composed by a variable number of experts (from 20 to 60, depending on the expected number of products – the experts were selected by ANVUR from list of applicants according to their publication records and their area of expertise). In turn, these experts relied onto 14,500 external peer reviewers, working in domestic and foreign institutions;
- in the last two exercises, the evaluating agency (ANVUR) requested to the experts a preassigned distribution of journals, according to the world distribution of impact. As a consequence, the top list of journals should correspond to the best 10% of the world production; nevertheless, more than 30% of the submitted products to the last exercise ended up in this category (because the exercise considers only the best products);
- depending on the research area, two assessment procedures have been followed:
  - *bibliometric assessment* consisted of combining the ranking of the journal according to the Impact Factor and the citations obtained by a specific article – articles in highly ranked journal with limited citations and/or highly quoted articles published in low ranked journals were peer reviewed;
  - *peer review assessment* consisted of a product being separately assigned to two experts, who independently selected an external peer reviewer; once the reviews were returned, a consensus report was drafted by the experts. In case of significant disagreement, a third reviewer was introduced, and the final assessment has to be approved by coordinator of the group of experts.

In both cases the submission to experts were non-blind, and the evaluators may have formed their opinion looking at the place of publication, in what has been called as "informed peer review".

2. The impact of the research assessment

The evaluation of the product is normalised according to the means in each research area, leading to an indicator which combines quality and quantity assessment of a research field in a university.[1] This indicator counts for three-fourths of the funds allocation, and is then

[1] From a technical point of view, the indicator consists of the share of scores attained by a single university/department over the total scores achieved at the national level by all institutions. That share is then applied to the distribution of funds. If a university/department performs above the average, it will obtain a funding share which exceeds the corresponding share computed on the personnel heads.

complemented with other indicators (PhDs, foreign students, external funding) in order to achieve the summary indicator to be applied to a funding scheme for universities. The most recent exercise lead to the distribution of ½ of total funding to public universities in Italy (1.4 billions of euro for 2016). Approximately 15% of total funding relies on the proper evaluation of research products.[2]

As such Italy belongs to *evaluation-based* systems (with the UK, Australia, New Zealand), to be contrasted with *indicator-based* systems (Norway, Denmark, Czech Republic). However, the 5-year interval is long enough to call for alternative methods of evaluation in the intermediate years. In addition, the results of the evaluation have trickled-down, directly or indirectly, to many other dimensions of the life of university departments. Many universities have used the scores obtained by their departments in the internal allocation of funds and promotions; the current accreditation of PhD programs is based on the research assessment of the teaching staff; newspapers articles have widely disseminated the results of the research assessment with reference to local universities, in order to drive the choices of students and their families.

Even if they are formally independent, the process of selecting new academics has been significantly influenced by the research assessment exercises. Selection in hard science research fields makes large use of bibliometric methods, while in soft science journal rankings have been adopted. Though I would not dare claiming that the introduction of assessment exercises has raised the standards of hiring in most disciplines, as a matter of fact in the most recent VQR the average score of newly hired/promoted researchers is higher than the average of permanent members (the indicator called IRAS2). This implies that new entrants in the academia have introjected the assessment approach in shaping the way in which they publish their research outputs.

While the VQR asks for the assessment of "originality, relevance, exposure to international debate", what is more perceivable (and perceived) is the internationalisation of the domestic production. Publishing in a foreign language (notably in English) has become the dominant strategy in several fields. As a consequence, many Italian journals which used to publish in Italian opted for the English language. A related issue is the multiplication of the number of papers via the diffusion of co-authorship. Since the VQR rules allow for the same product being submitted by more than one author (as long as they belong to different research entities), many authors have followed a strategy of risk diversification, by developing joint research projects in the expectation that at least one of them would obtain publication in a highly ranked journal.

3. The recent VQR (2011-14)

The most recent research assessment exercise ended in February 2017, with the official presentation of global report on the Italian research activity accompanied by specific reports for each research areas and for the social impact activity. 96 Universities participated to the exercise, together with 12 PRO's (Public Research Organisations) and 26 other institutions on a voluntary basis. The distribution of 118,036 products received for evalua-

[2] To be honest, the impact on funding is less dramatic in the short run, because of high persistence on historical values: each university cannot receive ±2% of what it has received the previous year, thus strongly attenuating whatever result could obtain from the research assessment.

tion is reported in Table 2, where one can easily detect few regularities. Compliance rates vary across research areas, oscillating between 90% and 97%.[3] Journal articles represent the dominant submission for hard sciences (reaching 98% in Biology and Medicine), while collected papers (edited volumes) prevail in the social sciences and humanities. Books have almost not been submitted in bibliometric areas, while they represent one fourth of all submissions in some non-bibliometric areas. The residual category [including musical compositions, designs, projects (architecture), performances, exhibitions, arts objects, databases and software] account for a small fraction of the total output submitted to the assessment. This does not produce a representative snapshot of the research activity of universities (PROs have similar composition), because the limit imposed to two products per researcher. Rather it allows monitoring of what can be considered as relevant scientific productivity of the entire research community.[4]

Table 2. Distribution of products by research area and type of output – Italy VQR 2011-14

Research area	Number of products expected	Number of products submitted	% missing	Journal articles	% on total by area	Books	% on total by area	Articles in collected paper volumes	% on total by area	Patents	% on total by area	Other products	% on total by area
1	6,680	6,007	10.1	5,278	87.9	66	1.3	632	10.5	1	0.0	17	0.3
2	10,923	10,562	3.3	10,244	97.0	22	0.2	154	1.5	18	0.2	121	1.1
3	7,232	7,033	2.8	6,916	98.3	18	0.3	75	1.1	23	0.3	1	0.0
4	4,638	4,398	5.2	4,052	92.1	38	0.9	276	6.3	2	0.0	28	0.6
5	11,706	11,033	5.7	10,831	98.2	18	0.2	154	1.4	13	0.1	15	0.1
6	18,148	16,936	6.7	16,612	98.1	54	0.4	256	1.5	7	0.0	1	0.0
7	7,849	7,470	4.8	6,848	91.7	59	0.9	535	7.2	14	0.2	9	0.1
8a	3,659	3,468	5.2	924	26.6	758	23.4	1,633	47.1	11	0.3	87	2.5
8b	3,010	2,818	6.4	2,497	88.6	12	0.6	299	10.6	5	0.2	1	0.0
9	12,074	11,339	6.1	10,055	88.7	67	0.6	1,146	10.1	45	0.4	21	0.2
10	9,363	8,761	6.4	2,822	32.2	1,944	23.9	3,819	43.6	0	0.0	30	0.3
11a	6,476	6,122	5.5	2,098	34.3	1,661	28.6	2,265	37.0	0	0.0	6	0.1
11b	2,385	2,292	3.9	2,026	88.4	80	3.8	175	7.6	0	0.0	4	0.2
12	8,973	8,502	5.2	3,373	39.7	1,977	26.3	2,893	34.0	0	0.0	2	0.0
13	9,039	8,300	8.2	6,041	72.8	638	8.7	1,490	18.0	2	0.0	47	0.6
14	3,242	2,995	7.6	1,232	41.1	699	24.6	1,024	34.2	0	0.0	2	0.1
Total	125,397	118,036	5.9	91,849	77.8	8,111	7.5	16,826	14.3	141	0.1	392	0.3

Note: The evaluation is organised in the following research areas: Mathematics (Area 1) - Physics (Area 2) - Chemistry (Area 3) - Earth Sciences (Area 4) - Biology (Area 5) - Medicine (Area 6) - Agricultural and Veterinarian Sciences (Area 7) - Civic Engineering (Area 8a) - Architecture (Area 8b) - Industrial and communication Engineering (Area 9) - Humanities (Area 10) - History and Philosophy (Area 11a) - Psychology (Area 11b) - Law (Area 12) - Economics and Statistics (Area 13) - Social Sciences (Area 14).

The assessment of each product was conducted according to three criteria:

1. *Originality*, to be intended as the degree according to which the publication is able to introduce a new way of thinking about the object of the research;
2. *Methodological accuracy*, to be intended as the degree according to which the publication adopts an appropriate methodology and is able to present its results to peers;
3. *Actual or potential impact*, to be intended as the level of influence – current or potential – that the research exerts on the relevant scientific community.

[3] It is important to recall that a protest organised in some universities led a fraction of university professors to refuse to submit their required output. However, in the first VQR, the submission rate for universities was 95.09% of the expected output, while it went down to 933.82 during the second one.

[4] The rules prevented the submission of textbooks, working papers and self-publications.

Each publication was attributed a quality profile:

- **Excellent** (weight 1) if it falls in the top decile of the world distribution of publications in the research area;
- **Good** (weight 0.7) if it falls in the 70-90% segment of the distribution;
- **Fair** (weight 0.4) if it falls in the 50-70% segment of the international distribution;
- **Acceptable** (weight 0.1) if it falls in the 20-50% segment of the distribution;
- **Limited** (weight 0) if it belongs to the 0-20% lowest segment of the distribution;
- **Impossible to evaluate** (weight 0) was assigned to missing publications or publications that were impossible to evaluate.

As one can easily expect, any evaluation of a product following the above-mentioned criteria contains some degree of arbitrariness. One can initially consider the language of publication as a proxy for the exposure to the international debate. An inspection to Table 3 seems to suggest that what are considered as bibliometric sectors (in light grey) are largely open to the international debate. From this perspective, the research area 13 (Economics and statistics) could be considered equally open to internationalisation. These areas have mostly relied on automatic assignment of products to the evaluation categories, using the principle that journal with high impact factors are generally speaking more selective in acceptance, and therefore impose higher standards of quality. This principle is complemented with the use of papers' citations, which should capture the relevance of the contents for the scientific debate.

The evaluation in non-bibliometric areas relied on peer review (with the exception of the research area 13, which adopted a ranking of the journals based on the impact factors). If the replacement of an algorithm with human reviewers may be welcome in terms of adherence to the suggested evaluation principles, it introduces the problem of potential disagreement among the reviewers, which is likely to motivate the lower fraction of "excellent" and "good" evaluation recorded in the non-bibliometric areas (see Table 4).

Table 3. Language of the products submitted to VQR 2011-14

Research area	Italian	English	Any other foreign language	Information n/a	Total
1	142	5,907	12	1	6,062
2	69	10,514	3	2	10,588
3	41	6,850	6	0	6,897
4	157	4,265	8	0	4,430
5	107	10,858	21	0	10,986
6	506	16,145	42	0	16,693
7	534	6,998	9	0	7,541
8a	2,239	1,171	46	0	3,456
8b	137	2,692	3	0	2,832
9	154	11,401	9	0	11,564
10	5,574	2,051	1,119	0	8,744
11a	4,295	1,454	374	0	6,123
11b	301	1,962	13	0	2,276
12	7,671	678	139	0	8,488
13	1,897	6,451	35	2	8,385
14	1,866	985	120	0	2,971
Total	**25,690**	**90,382**	**1,959**	**5**	**118,036**

Table 4. Distribution of products by research area and received evaluation – VQR 2011-14

Research Area	A (excellent)	B (good)	A+B	C (fair)	D (acceptable)	E (limited)	F (non classified)	Total
1	38.4	28.0	*66.4*	18.2	10.8	4.2	0.4	100
2	62.2	21.6	*83.8*	10.4	4.7	0.9	0.1	100
3	49.2	32.0	*81.1*	12.9	4.6	0.8	0.6	100
4	27.9	29.7	*57.5*	21.6	14.2	5.5	1.2	100
5	37.3	31.5	*68.7*	19.0	9.3	1.8	1.2	100
6	39.5	25.8	*65.3*	17.8	11.9	3.6	1.4	100
7	28.4	31.5	*59.8*	19.4	14.9	5.4	0.5	100
8a	8.6	34.2	*42.8*	35.9	16.0	4.8	0.5	100
8b	37.6	29.3	*66.9*	17.7	12.6	2.5	0.2	100
9	38.6	27.6	*66.2*	18.2	12.3	2.7	0.7	100
10	18.1	46.2	*64.3*	25.4	8.7	1.4	0.2	100
11a	16.1	42.4	*58.5*	29.2	10.2	1.8	0.3	100
11b	30.8	23.4	*54.2*	19.1	18.7	6.8	1.2	100
12	7.8	41.2	*49.0*	35.9	12.2	2.2	0.7	100
13	24.6	22.9	*47.5*	17.9	19.5	12.7	2.3	100
14	8.3	32.5	*40.9*	34.5	20.0	4.4	0.3	100
Total	**32.6**	**30.8**	*63.4*	**20.7**	**11.6**	**3.5**	**0.8**	**100**

4. The receipt of the research assessment exercises in the academic community

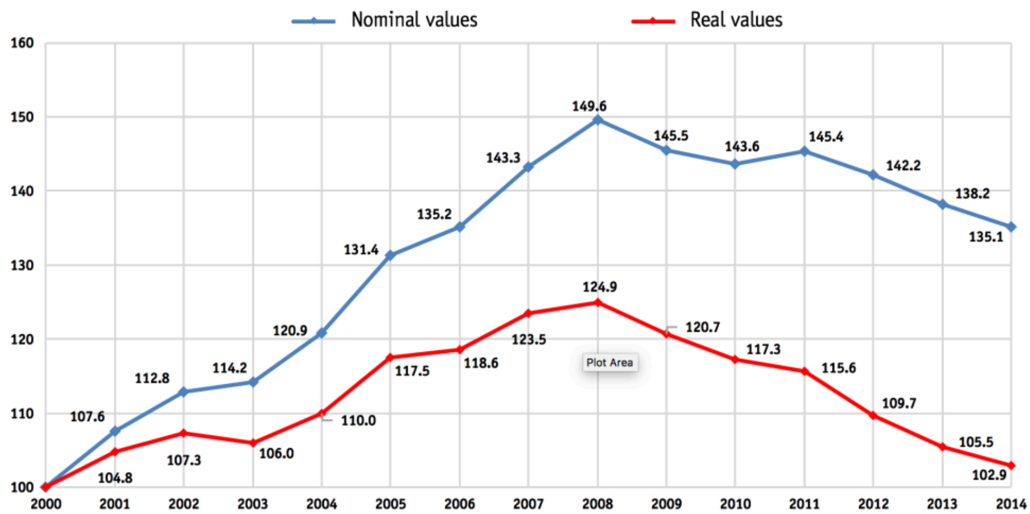
These exercises have generated enthusiasm and collaboration as well as suspicion and resistance. A large fraction of academics definitively cooperated with the exercise, organising the submission within each department and accepting to review the product. A smaller fraction opposed it, on the arguments that these exercises were misleading the Italian research towards irrelevant topics, were promoting harmful competition among research agencies and were destroying the weakest segment of the academia (very often located in Southern universities).[5]

My impression is that the main argument against the research assessment exercise runs as follows: "the assessment legitimizes budget cuts, especially against southern universities. If we want to save the equal opportunity in accessing universities, we should oppose any assessment which associate funding and results". This argument has some plausibility, especially when looking at Figure 1, which shows the trends in state funding to Italian universities in nominal and real (i.e. deflated by the price variation) terms. Remember that the first exercise with impact on funding was launched in 2011, when the decline in resources became more pronounced. Although the actual impact was not disruptive (due to safeguard clauses – see above), the linkage of resources to assessment opened the risk of "poverty traps": a poorly performing university received fewer resources and was therefore less likely to improve its performance in the next round of assessment. Budget cuts curtailed hiring possibilities, which were only later released in

[5] Perhaps the most representative instances of this aversion towards evaluation performed by ANVUR can be traced in the following websites (unfortunately, all in Italian): www.roars.it (usually covering topics related to assessment methods); http://www.flcgil.it/universita/ (the website of the main union of university workers); http://firmiamodimissionianvur.org/ (more than 2,000 researchers signed a petition asking for the dismissal of the board of ANVUR, the evaluation agency).

correlation with performance. Thus poorly performing universities were supposedly prevented from hiring better researchers in order to revert their rank position.

Figure 1. Total public revenues accrued to Italian public universities (2000=100)



Despite its simplicity, this line of argument is substantially flawed. During the first decade of the present century, the hiring procedures of Italian universities were reformed, moving from a format of centralised competition to one of local competitions. Each department was left almost free to hire or promote whoever they deemed worthy to be hired. The first exercise (VTR) did not provide a clear picture of the average performance, because it was designed to assess excellence within each university, without considering who wrote what. The second exercise (VQR 2004-10) for the first time revealed that a non-negligible fraction of researchers was unable to submit any research product at all. The third exercise (VQR 2011-14) provided evidence of some convergence of universities towards the mean, thanks to the change in the grading procedure (missing submissions were no longer penalised with a negative grade) but also to the injection of new resources that made possible to all universities the hiring of new scholars.

5. Open issues for future assessment exercises

In the immediate aftermath of the publication of the results of the third exercise, several suggestions have emerged in the press as well as in official forums. Some of them were mainly technical, some other more philosophical. In the following I will review them in brief.

The first concerns the potential bias contained in the evaluation. Given existing rules, co-authored papers to be submitted to foreign journals have the highest probability to receive a high grade. This implicitly “delegates” to foreign editors (and publishers) the choice of what is to be considered relevant for the international debates. Topics that are outside the mainstream, or that are simply concerned with national debates, are likely to appear at best in local journals, which then receive lower evaluation even by referees. Still, most of Italian journals do not yet have standard double blind reviewing procedures, inducing the suspicion that the quality of their articles may be lower.

The absence of domestic databases on publications and citations makes it impossible to introduce a dual layer system, where articles and books in Italian could gain more visibility. The use of peer reviewers is not a panacea, for various reasons. Especially in the social sciences, where the ideological content of the arguments is important, the judgment of the reviewer may be biased by strategical concerns (by attributing a lower score to an author, one may be tempted to alter the competition among different schools of thought). In addition, peer review of papers that have already undergone a real blind review process represent in inherent contradiction: suppose that the final reviewer spots an evident error; who has to be blamed, the author, the journal referees, or the editor of the journal? Finally, the peer review is expensive. Consider the following back of the envelope calculation: in the most recent exercise 52,060 products (corresponding to 44.1% of total production) underwent a double review; each reviewer received 30 euro per review, leading to a total cost above 3 million euro, which is a cost that cannot be frequently afforded.

The second aspect concerns the different publication strategies of different research communities. On average applied physics scholars publish more than 30 papers per year, because the number of co-authors can easily exceed one hundred. The corresponding figure for a theorist in mathematics may not reach one paper per year. To partially account for these differences the scores are normalised by research area, but this does not reduce the evident advantage of sectors where the scholar may select their best production from a larger set of papers.

A related issue deals with the weighing of different products. The most recent exercise introduced for the first time a different weighing for books vis a vis journal articles: under specific request of the author, a book could have been considered as equivalent to two articles, thus satisfying the requirement of submission. But the principle could be extended to other categories of products, because an article collected in a book is probably subject to less scrutiny than an article in a journal. Articles and/or books could be weighed by the number of co-authors. And so on.

A further issue that has been raised deals with the boundaries of research areas. So far the assessment exercises have considered aggregation of research fields (*settori scientifico-disciplinari*) under which academics have been hired to teach. This does not have any correspondence to other classification criteria (like ERC) and tend to penalise cross-disciplinary research. In principle, nothing prevents redesigning of the evaluation areas, but this interferes with the academic careers, which represents the strongest incentive to publish (at least for academics). Thus, a net separation between research assessment and promotion criteria would be required before addressing this problem.

A final point deals with the potential trade-off between teaching and research. The assessment is conducted without any reference to the resources available/invested in research, including the time absorbed by teaching. Most universities in peripheral areas lament the excess burden of teaching created by the chronic lack of staff. From an intuitive point of view, a proper assessment should correct for differences in the starting conditions. Otherwise stressing research results as unique measure for scholars' quality is detrimental to the effectiveness of teaching, because scholars will devote their best energies to article writing. There are possible solutions to avoid this trade-off: if each academic could choose over a menu of different combinations of teaching loads and commitment to publications, we could observe a possible sorting of scholars according to their preferences and abilities. This would require a revision of the procedure of assessment, because scholars should then be weighed or converted into full-time equivalents.

Overall, the unsolved issue for the Italian research assessment exercises seems to be whether the results should be interpreted as *monitoring the system* (in order to ensure accountability vis-à-vis the tax-payers) or rather a *research quality assessment* (intended to promote excellence). The Ministry of Education oscillates among these two interpretations, which however lead to alternative policy suggestions. According to the former perspective, uniformity of performance is a goal, and the weakest universities should be sustained in order to grant a common standard of tertiary education across the country. According to the latter, the best universities/departments should obtain even greater resources, given their good evaluations obtained in the assessment.