Scholarly Assessment Reports

Bonaccorsi, A. (2020). Two Decades of Experience in Research Assessment in Italy. *Scholarly Assessment Reports*, 2(1): 16. DOI: https://doi.org/10.29024/sar.27

RESEARCH

# Two Decades of Experience in Research Assessment in Italy

Andrea Bonaccorsi
Univ Pisa, IT
a.bonaccorsi@gmail.com

The paper reviews the Italian experience in the evaluation of research in the 2000–2020 period. The initial exercise (VTR 2000–2003) did not involve all researchers and had no impact on funding. After a long political and cultural debate there was a decision to create an independent Agency in charge of a periodic research assessment, involving all researchers, and having impact on performance-based funding. The legislation was approved in 2006 and the Agency was created in 2010–2011. In parallel, a major reform of academic promotion was approved in 2010. The Agency (ANVUR) launched three exercises, two of which have been completed and published (*Valutazione della Qualità della Ricerca*, or Assessment of Research Quality, VQR 2004–2010 and VQR 2011–2014). It also developed a complete array of quantitative indicators to be used as a threshold for candidates to the academic promotion (Habilitation). The paper offers detailed evidence of the evaluative framework, the main methodological and practical problems and the changes and adaptations introduced over time. It concludes with several policy implications.

**Keywords:** research assessment; bibliometrics; evaluative framework; Italy; ANVUR; VQR

## 1. Introduction and caveat

This paper is the first of a two-part essay in which I try to offer a complete and critical view of the Italian experience of research assessment. Part I is dedicated to the detailed description of the experience, while in Part II I will try to make justice of the criticisms and controversies generated by research assessment.

Italy is an interesting case study for the international community working on science policy and research evaluation, on the one hand, and on informetrics and bibliometrics, on the other hand. It is the only large Continental European country in which research assessment has been made mandatory, has implications on university funding, and is carried out on a large scale at regular intervals. With more than 180,000 research products evaluated, the VQR 2004–2010 was the largest institutional exercise ever carried out. With more than 40,000 titles and 15,000 journals rated, the journal rating system is one of the largest available and has survived the criticisms that in other countries, such as France and Australia, led to its cancelation.

Another intriguing reason of interest is that in the Italian context the evaluative informetrics, in particular the use of bibliometric indicators, has been introduced suddenly and rapidly, generating in a few years a lot of controversies, but also large opportunities for institutional learning and adaptation.

There is an important caveat to my analysis: I have been a member of the Board of the Italian Agency (ANVUR) during the startup phase (2011–2015) and I have been personally responsible for some of the procedures, and collectively responsible for all decisions made in that period. In the current and the companion paper I will try to examine the experience in a professional way, by using the available evidence systematically and balancing the arguments. The reader will evaluate whether my account is worth of attention.

In this paper I first describe the events and decisions that led to the various research assessment exercises (Section 2). I then work backward, from the legislation and the administration to the main principles, objectives, purposes and criteria of the evaluative framework (Section 3), as reconstructed. The reasons for this inversion of the logical flow (not from principles to execution but the other way round) will be clear to the reader only after reading these sections. Section 4 discusses the reception of the research assessment in the university landscape and Section 5 enlarges the description to another assessment activity carried out by the Agency in the context of National Scientific Habilitation of candidates to the academic career.

In the companion paper I will examine all criticisms that have been raised against the research assessment in the peer-reviewed literature and will try to balance the various arguments. At the end I will try to propose a balanced judgment of the overall experience.

## 2. Research assessment in Italy. A long tale

The current experience of research assessment goes back to the year 2000 and covers three completed exercises and a work-in-progress. I summarize most technical elements of the four exercises (from VTR 2000–2003 to VQR 2015–2019) over two decades in **Table 1**. The discussion of institutional, methodological and political issues is carried out in the text.

### *2.1. The first experiment: VTR 2000–2003*

As in other advanced countries, Italy experienced a process of shift from a centralized model of university administration to a model based on autonomy, although the process started only in the '90s. After granting organizational and financial autonomy to universities there was general agreement on the need to provide the higher education system with advanced instruments for steering, accountability and responsibility.

This orientation led to the creation of CIVR (*Comitato di Indirizzo per la Valutazione della Ricerca*, or Research Assessment Committee), a non-permanent body that took the responsibility for research assessment. CIVR launched the VTR (*Valutazione Triennale della Ricerca*, or Three-year Research Assessment), covering the year 2000–2003. The exercise was entirely based on peer review and the selection of products was done by departments (Cuccurullo, 2006).

According to many observers, the main limitations of the VTR were the procedure for the selection of products and the lack of impact on funding. The selection of products to be submitted was the responsibility of departments, which in theory should have followed criteria of research quality. As a matter of fact, given the lack of experience in research assessment, but also given the lack of consequences of inappropriate choices, products submitted were selected using a variety of criteria, often not correlated to the quality of products but to academic rank or other considerations. Abramo, D'Angelo and Caprasecca (2009) examined the selection of products by universities and concluded that they did not identify their best publications in terms of citation impact.

The initial experience led to a large debate, which identified several policy priorities for the future. First, the evaluation of research should become a permanent activity, allocated to a professional structure with a clear long-term mandate. Second, research assessment should have an impact on funding of universities. Third, in order to create a diffused evaluation culture, all researchers should be submitted to evaluation and should be responsible for the selection of their products. In terms of methodological choices of VTR several contributions pointed to the need to enlarge the assessment toolbox in order to include bibliometrics (Abramo, D'Angelo and Caprasecca, 2009; Franceschet and Costantini, 2011). In addition, there was a keen awareness of the need to join the European framework for quality assurance in higher education. After the implementation of the Bologna process, in fact, Italy had not yet met the formal requirements for joining ENQA, the European institution devoted to the quality assurance of educational activities, an essential step in the mutual recognition of curricula among European countries.

### *2.2. The creation of the new Agency 2006–2011*

Given these orientations, a long parliamentary debate led to the creation, in 2006, of a national agency, called ANVUR (*Agenzia Nazionale per la Valutazione del sistema Universitario e della Ricerca*, or National Agency for the Evaluation of Universities and Research institutes).[1] The law was explicit in asking that the results of evaluation should have an impact on public funding.[2] The great expectations on the role of the Agency led to a legislative design with a complex implementation. To start with, there was a need for a Ministerial decree for the creation of the structure and its funding. The responsible structure was the *Ministero dell'Istruzione, Università e Ricerca* (MIUR, Ministry of Education, Universities and Research). Although there was the suggestion to create an independent Authority, fully separated from the Ministry in terms of annual budget, this solution was not pursued due to the complexity of the political decision (particularly after a period in which several Authorities had been created in other fields of regulation). The regulation of the Agency took a long period and was established only in 2010.[3]

The legislation designed a complex procedure for the nomination of the members of the Board (*Consiglio Direttivo*), with the aim to give it independence and authoritativeness. An independent search committee was nominated, with members designated from OECD, the European Research Council, Accademia dei Lincei and the National Student Council (*Consiglio Nazionale degli Studenti*). The search committee evaluated self-candidatures and selected a list of 15 candidates. The Minister selected the seven members of the Board within the list. The selected members were to be approved by the entire Government (*Consiglio dei Ministri*). After government approval, the members had to be approved by two Culture Commissions of the branches of the Parliament, the Senate and the House of Representatives. Finally, their nomination was to be signed by the President of the Republic.[4]

---

[1] Legge 24 novembre 2006, n. 286. All documentation on the legislation and the composition of the Board is available at https://www.anvur.it/anvur/riferimenti-normativi/.

[2] "I risultati delle attività di valutazione dell'ANVUR costituiscono criterio di riferimento per l'allocazione dei finanziamenti statali alle università e agli enti di ricerca" (The results of the evaluation activities of ANVUR constitute reference criteria for the allocation of government funding to universities and PROs). Legge 24 novembre 2006, no. 286, art. 2, comma 139.

[3] Decreto del Presidente della Repubblica 1 febbraio 2010, n.76.

[4] Decreto del Presidente della Repubblica 22 febbraio 2011.

**Table 1:** Synopsis of main features of Research asssessment exercises in Italy. Year 2000–2020.

| | VTR | VQR I | VQR II | VQR III |
|---|---|---|---|---|
| Period covered | 2000–2003 | 2004–2010 | 2011–2014 | 2015–2019 |
| Year of start | 2003 | 2011 | 2015 | 2020[a] |
| Year of publication | 2004 | 2013 | 2017 | 2022 |
| Organization | CIVR | ANVUR | ANVUR | ANVUR |
| Subject evaluated | 77 universities<br>12 PROs (research institutions under the responsibility of MIUR)<br>13 public and private research institutions | 96 universities<br>12 PROs (research institutions under the responsibility of MIUR)<br>26 public and private research institutions | 96 universities<br>12 PROs (research institutions under the responsibility of MIUR)<br>27 public and private research institutionr | n.a. |
| Evaluation method | Peer review | Peer review + Bibliometrics | Peer review + Bibliometrics | Informed peer review |
| Bibliometric indicator | None | Normalized number of citations until 2011 + Journal impact factor | Normalized number of citations until 2015 + Journal impact factor<br>Journal indicators<br>- 5 Years Impact Factor<br>- Article Influence Score (AIS) for WoS<br>- Scimago Journal Rank (SJR)<br>- Impact per Publication (IPP) for Scopus | To be decided by GEVs. |
| Bibliometric source | No | WoS, Scopus, MathSciNet | WoS, Scopus, MathSciNet | To be decided by GEVs. |
| Submission decision | Department | University or PRO based on individual proposals for submission by researchers (n = 61822) | University or PRO based on individual proposals for submission by researchers (n = 52677) | University or PRO based on individual proposals for submission by researchers |
| Type of products | Journal article<br>Book<br>Book chapter<br>Proceedings of national and international conference<br>Patent<br>Design<br>Performance<br>Exhibition<br>Manufacture<br>Art opera | Journal article<br>Book<br>Book chapter<br>Conference proceedings<br>Critical review<br>Commentary<br>Book translation<br>Patent<br>Prototype<br>Project plan<br>Software<br>Database<br>Exhibition<br>Work of art<br>Composition<br>Thematic paper | Same as 2004–2010 | Same as 2004–2010 |

(Contd.)

| | VTR | VQR I | VQR II | VQR III |
|---|---|---|---|---|
| Number of products per capita | At least one product each 4 researchers (universities) or 2 researchers (PROs) | 3 products per university staff<br>6 products per PRO staff | 2 products per university staff<br>4 products per PRO staff | 2 products per university staff<br>4 products per PRO staff<br>(with possibility of compensation at university level) |
| Total number of products evaluated | 17329 evaluated<br>18500 submitted | 184878 evaluated | 118036 evaluated | n.a. |
| Expert panel | 20 expert panels | 14 GEV (Gruppi di esperti della valutazione)<br>450 members | 16 GEV (Gruppi di esperti della valutazione)<br>436 members | 17 GEV (Gruppi di esperti della valutazione) + GEV Third mission |
| Choice of expert | Call + nomination CIVR | List of experts from previous call (n > 3000) + nomination ANVUR | Public call + nomination ANVUR | Public call + random extraction with quotas |
| Number of referees | 6661 referees of which 1465 from abroad | >14,000 | >13,000 | n.a. |
| Main quality criteria | | Relevance to the field<br>Novelty<br>Internationalization | Originality<br>Methodological rigor<br>Attested or potential impact | Originality<br>Methodological rigour<br>Impact |
| Classes of merit | Excellent (top 20%)<br>Good (60–80%)<br>Acceptable (40–60%)<br>Limited (bottom 40%)<br>Scale of values shared by the international community | Excellent (top 20%)<br>Good (60–80%)<br>Acceptable (50–60%)<br>Limited (bottom 50%)<br>Scale of values shared by the international community | Excellent (top 10%)<br>High (10–30%)<br>Fair (30–50%)<br>Acceptable (50–80%)<br>Limited (bottom 20%) | A. Excellent and extremely relevant<br>B. Excellent<br>C. International relevance<br>D. National relevance<br>E. Limited or no relevance |
| Score | Excellent 1<br>Good 0.8<br>Acceptable 0.6<br>Limited 0.2 | Excellent 1<br>Good 0.8<br>Acceptable 0.5<br>Limited 0 | Excellent 1<br>High 0.7<br>Fair 0.4<br>Acceptable 0.1<br>Limited 0 | Not defined in the Ministerial decree |
| Penalty | Not applicable | Proven cases of plagiarism or fraud (–2)<br>Product types not admitted by the GEV, or lack of relevant documentation, or produced outside the 2004–2010 period (–1)<br>For failure to submit the requested number of products –0.5 for each missing product | Not assessable 0 | Not assessable 0 |
| Aggregation | Scientific structure<br>Mega (>74 products)<br>Large (25–74)<br>Medium (10–24)<br>Small (less than 10) | Department<br>Scientific area<br>University | Department<br>Scientific area<br>University | Department<br>Scientific area<br>University |

(Contd.)

| | VTR | VQR I | VQR II | VQR III |
|---|---|---|---|---|
| Cost | 3,55 million euro | 10,570 million euro (including CINECA costs) | 14,7 million euro (including costs at university level) | n.a. |
| Impact on funding | No[(b)] | 13% of total funding (2013)<br>16% (2014) | Appr. 25% total funding (1,5 bn euro) | n.a. |
| Additional information | Human resources<br>International mobility of researchers<br>Funding for research projects<br>Patents<br>Spinoff companies<br>Research contract | Overall indicator IRFS1<br>50% based on VQR<br>50% based on the following information<br>Amount of external research funding (10%)<br>Quality of new recruits and promotion (10%)<br>Internationalization (10%)<br>Number of doctoral students and postdocs (10%)<br>Propensity to finance projects with endowment funds (5%)<br>Performance improvement compared to the VTR 2000–2003 evaluation (5%) | Same as 2004–2010 | n.a. |

Legenda

(a) The start of the process was postponed due to the Covid-19 crisis.
(b) As a matter of fact, the Ministry used the VTR data to allocate a small share of funding (2%) in 2009. The move was criticized for using obsolete data, backing to the start of the decade. This criticism accelerated the pressure for launching a new exercise and using new data for allocation of funding.

This complex selection and nomination procedure was designed to build up the independence of the Agency from the Ministry. The lack of formal separation from the budget of the Ministry, however, made this provision weaker, insofar as the allocation of resources was dependent on budgetary decisions of a political type. This financial subordination was solved only later on (see below). At the same time the nomination procedure was aimed at giving the Agency an authoritative and legitimate role, based on professionalism, independence and transparency.

With such a political and administrative complexity, the Agency, whose legislative creation dates back to 2006, was created in 2010 and put into action only in May 2011. This may explain why the Agency started immediately to work on research assessment, without a long consultation with the scientific communities, universities, and public research organizations (PROs). This can be considered a violation of the recommendations that are usually formulated for a sound research assessment process, that is, a large discussion with the academic community *before* starting operational activities.

### *2.3. The first large scale research assessment. VQR 2004–2010*

A Ministerial decree dictated the main objectives of the research assessment exercise but also defined several technical details.[5] The procedure started in July 2011 with the publication of the general criteria by the Ministry.[6] ANVUR nominated the GEVs (*Gruppi di Esperti della Valutazione*, or Evaluation Expert Panels). In order to save time, members of the GEV were selected among the list of self-candidatures that was elicited by CIVR following the previous research assessment exercise (VTR). The activites of CIVR were interrupted due to the opening of the parliamentary process that eventually led to the new agency. The GEVs were created following the administrative definition of 14 scientific areas, defined as aggregations of disciplines called *Settori Scientifico-Disciplinari* (SSD, or Scientific Disciplinary Fields), which is a long standing feature of the Italian system.[7] These areas are defined by the National University Council, or CUN (*Consiglio Universitario Nazionale*). The highest level of aggregation is formed by 14 broad disciplinary areas, labelled *Area CUN*. ANVUR also nominated the Presidents of GEVs, trying to involve outstanding scientific authorities in the respective fields.

A clear divide was put in place by expert panels between those fields in which bibliometric criteria were adopted (STEM disciplines, which fall in CUN areas 1–9) and those in which peer review was adopted (SSH, in CUN areas 10–14). Remarkable exceptions were Architecture (which adopted peer review although it is included in the Engineering area) and Psychology (which adopted bibliometrics although it is included in Humanities). Another exception was Economics, illustrated below.

In order to put this approach into context a number of remarks should be made (see also Ancaiani et al. 2015). First, there was a mandate in the legislative text to make "prevailing" use of peer review. In the parliamentary debate there had even been opinions that wanted peer review to be the exclusive methodology, as was done in VTR. This mandate was interpreted by ANVUR as making peer review representing more than 50% of evaluations. In other words, apart from methodological considerations, the budget constraint of the VQR made it impossible to rely exclusively on peer review. Second, peer review was also adopted in STEM disciplines for those articles that were too young to have received citations, or for which the citation window was too short to make citation analysis reliable. Scores from peer review were then aggregated with scores from bibliometrics, an approach which attracted some criticism, as we will see in the companion paper. Third, there was an important exception, which was implemented in the area of Economics and Statistics (CUN area 13). While this area is part of Social Sciences and Humanities (SSH), the panel decided to evaluate books and other non-article items via peer review, but to assess all articles via bibliometrics. Bibliometric indicators for non-indexed journals were estimated using a regression model, calibrated on indexed journals, and assuming the number of Google Scholar citations as independent variable. The panel then made an experiment to compare peer review and bibliometric indicator, which elicited a discussion, which we review in the companion paper.

In those GEVs in which bibliometric analysis was implemented, two indicators were used: the normalized number of citations of the individual article, and the journal impact factor, or equivalent. The GEVs had the mandate to publish evaluation criteria that were tailored to the scientific areas. In practice this meant two main decisions: the choice of the bibliometric database, and the algorithm for the aggregation of the indicators. While ANVUR left the panels free to choose, most panels adopted the Web of Science (WoS), while a few used several journal indicators both from WoS and Scopus. The issue of algorithm selection is more complex, and is at the origin of controversies, as we will see.

Let us illustrate the issue. The number of citations received by an article in the relevant citation window was normalized against the distribution of citations of all articles of the same year in the same Subject Category at world level. This

[5] The documentation on the three VQR exercises is available at https://www.anvur.it/attivita/vqr/.

[6] Decreto Ministeriale 15 luglio 2011 no. 17.

[7] While scientific fields (SSD) at more granular level are subject to periodic revision and update, following scientific developments, the main broad areas are relatively stable. They are labelled *Aree CUN*, following the name of the institution in charge of their definition, *Consiglio Universitario Nazionale* (National University Council), an elective body. All members of academic staff are affiliated to just one SSD. They are as follows: Area 1 (Mathematics and Computer Sciences); Area 2 (Physics); Area 3 (Chemistry); Area 4 (Earth Sciences); Area 5 (Biology); Area 6 (Medicine); Area 7 (Agricultural and Veterinary Sciences); Area 8 (Architecture and Civil Engineering); Area 9 (Industrial and Information Engineering); Area 10 (Ancient History, Philology, Literature and Art History); Area 11 (History, Philosophy, Pedagogy, Psychology); Area 12 (Law); Area 13 (Economics and Statistics); Area 14 (Political and Social Sciences).

resulted in a number which represents a percentile of the world distribution. The journal impact factor (whether WoS or Scopus) was normalized against the distribution of all journals in the Subject Category, again resulting in a percentile.

The technical issue was that the Ministerial decree mandated the classification of all products in ordinal categories (classes of merit), defining the range of the quantiles for the overall distribution of the score (see **Table 1** for the details). Products were then classified in one of the four classes according to the separate indicators. If the classifications were perfectly correlated (diagonal of the matrix) then the final class of merit was automatically assigned. If they were not correlated, the main decision was whether to give more weight to the citation or the journal indicators. Here the strategic decision of ANVUR was to leave the GEVs free to calibrate the two indicators according to the prevailing practice of the relevant scientific communities. This resulted in slightly different aggregation algorithms across the GEVs. The panels also sent a number of articles to peer review. This was done systematically for very recent articles and for those for which the two indicators delivered divergent results.

The work of GEVs produced a score for each of the products submitted. All Italian researchers have access to a personal page in a platform (loginmiur) managed by a large IT consortium of universities (CINECA) working on behalf of the Ministry. The scores were transmitted confidentially to all researchers via their personal page on loginmiur. This means that all researchers received a number for each of the products submitted. No individual information was disclosed. ANVUR made clear that the evaluation referred only to products and had a statistical meaning, hence it was not intended to evaluate researchers. The Agency explicitly warned universities against the possibility of asking researchers to disclose their personal scores on a voluntary basis. At the same time it is clear that the impact of research assessment in the Italian context has been very deep, and emotionally charged, given the personal engagement of all people.

After receiving the scores, ANVUR aggregated the indicators at three levels: (a) discipline; (b) department; (c) university. The aggregation at the level of scientific field or discipline (*Settore Scientifico Disciplinare*, SSD) was done if the number of researchers was at least four, in order to preserve the statistical secrecy. Since it may happen that researchers in the same field are affiliated to various departments, this level of aggregation gives universities a map of their internal quality at granular level. The aggregation at the level of department was done simply on the basis of the affiliation. At the level of universities the scores produced by GEVs were integrated with other indicators, which describe the overall research quality (see **Table 1**).

After integration of these indicators it was possible to produce several rankings of universities, using various weighting schemes. ANVUR prepared several variants on behalf of the Ministry, but then transmitted to the media the unique ranking that was selected *by the Ministry* as the basis for the allocation of the performance-based funding.

The results of the first VQR were presented in summer 2013 and obtained a large media coverage. They showed a large divide between universities in the South and those in Central and Northern regions. This opened a large debate on whether research assessment might produce unintended effects of deprivation for Southern universities, which may suffer from external diseconomies and lack of attractiveness due to the lower level of socio-economic development. At the same time, it was shown that Southern universities performed well in hard sciences, in which the scientific communities are international by nature, while the large gap in research quality was mostly concentrated in SSH and, in part, in Medicine. The debate went on. The law adopted a cautious approach to the allocation of performance-based funding, implementing a funding formula that prevented large reductions in overall funds as a direct consequence of the VQR (see section 4 for details).

The large media coverage of the VQR reflected the novelty of the exercise. The final ranking was unanticipated, with several surprising results, such as the relatively poor performance of many historical and large universities and of some established PROs, against the dynamism of a few medium-sized young universities.

### *2.4. Learning from experience. The second large scale research assessment VQR 2011–2014*

The second VQR started in 2015[8] and capitalized on the initial experience, introducing several improvements. The following modifications, some of which are discussed in the companion paper, were introduced:

- Two new groups of experts were created, separating Psychology from History, Philosophy and Pedagogy (since in Psychology it was decided to adopt bibliometrics) and Architecture from Civil Engineering (since in Architecture there was a need to adopt peer review for the evaluation of books, graphical works and projects).
- The classes of merit were modified, with a more uniform definition of the boundaries of the classes.
- The lowest score was no longer zero, a score which created a feeling of discomfort in many scholars.
- The algorithm for the aggregation of bibliometric indicators was modified by introducing a more general nonlinear transformation (see Anfossi et al. 2016 for details).

In addition, with a separate provision, the use of ranking for performance-based funding was modified, by shifting part of the allocation of funding from universities to departments, and selecting a number of excellent departments scattered across all universities.

[8] Decreto Ministeriale 27 giugno 2015 no. 458.

The results of the second VQR were anticipated in late 2016 at aggregate level and were published in detail in 2017. Overall they confirmed the North-South divide, although some convergence towards a higher average level of research quality was evident.

The new rules for the allocation of performance-based funding were implemented immediately, following the Budget law 2017 (Law 232/2016).[9] On the basis of VQR scores 350 excellent departments were identified and invited to submit a 5-year strategic plan for growth, covering the year 2018–2022. After the submission of plans, a national committee nominated by the Ministry of University and Research (MIUR)[10] selected the best proposals. There was a minimum number of departments per university to be selected (n = 1) and a maximum number (n = 15). The total number of departments selected was 180, with a budget of 271 million euro per year, totalling 1355 million euro over the five years. The best universities had several departments selected (Padova n = 13; Bologna n = 14), while 23 universities had just one. The minimum number was intended as a role model for all universities, including those that were found systematically at the bottom of the aggregate rankings. In economic terms the final allocation followed a lessicographic ordering. This initiative absorbed only part of the performance-based funding and created an important institutional innovation.

After the second VQR a legislative change was introduced, by establishing that the time window of the exercise should be five years.[11] The first VQR had to cover a longer period (2004–2010) in order to create continuity with the previous exercise (VTR 2000–2003). The second VQR covered four years (2011–2014), in the absence of a clear mandate. This legislative provision, included in the annual budget law, is also important because it not only dictated the time window for research assessment but also provided a mandatory procedure for budgeting. This made the research assessment exercise independent on the Ministry and its political decision making.

### *2.5. The third exercise and the Covid crisis*

Following this provision, the third VQR will cover the period 2015–2019. It has been launched in 2019 but postponed to the end of 2020 due to the Covid crisis. In the third exercise the following modifications have been introduced:

- The composition of GEVs must follow a set of rules, with a minimum number by categories (e.g. by academic rank, type of institution, or gender).
- The members of GEVs will not be nominated by ANVUR but extracted *randomly* from the list of candidates.
- All products will be peer-reviewed using an informed peer review methodology, associated in some cases to citation indicators.
- The number of products per researcher will be variable, allowing some researchers to compensate for others.
- The number of co-authors will be considered in the evaluation.
- Classes of merit do not have predefined quantitative levels of a score.
- A new GEV has been created, separating Management from Economics and Statistics.
- The exercise will also include the evaluation of third mission, with a dedicated GEV.[12]

In the companion paper I will examine carefully the main criticisms that have been raised against VQR. My preliminary statement is that several weaknesses have been addressed in moving from VQR 2004–2010 to the latest editions. On other points there is still significant discussion.

## 3. Reconstructing the evaluative framework

After the detailed description of the research assessment exercises carried out and in-progress, it is important to step back and ask whether there is an overall conceptual framework guiding the actions of the Agency.

In recent contributions Henk Moed has emphasized the need to place evaluative activities in an evaluative framework (Moed 2020a; 2020b). Research assessment has four interdependent dimensions: Policy and management, Evaluation, Analytics, and Data collection. The former two dimensions formulate the main issues underlying the research assessment, establish the objectives, and define a set of evaluation criteria. Analytics and Data collection, on the other hand, are the domain of evaluative informetrics, or the study of evaluative aspects of science and scholarship using informetric data and methodologies. Policy objectives and evaluation criteria cannot be founded or demonstrated informetrically but must be defined at higher levels by taking a non-neutral responsibility (ultimately a political responsibility).

In the Italian context, it is possible to reconstruct the overall evaluative framework by inspecting the legislative and ministerial documents that have shaped the research assessment activities. It must be emphasized that these documents define the operational aspects of the activities of the Agency and of the specific exercises without a formal and extended definition of the evaluative framework. Nevertheless, by reading available documents systematically and comparing them with the previous experience (VTR 2000–2003) it is possible to obtain a reasonably clear picture. I build

[9] The overall procedure is described in https://www.miur.gov.it/dipartimenti-di-eccellenza.

[10] The committe was formed by seven members: one nominated by the Prime Minister, two nominated by the Minister of University and Research, four nominated by the same Minister within a short list of candidates proposed by ANVUR and by the National Committee of Resarch Guarantors (Ministerial Decree n. 262, 11 May 2017).

[11] Legge 11 dicembre 2016, n. 232, art. 1, co. 339.

[12] The history of evaluation of third mission is very interesting but cannot be discussed here. See Blasi et al. (2018b) for a preliminary discussion and Blasi et al. (2019) for first empirical results on the complementarity between research quality and third mission indicators.

up the reconstruction following the items suggested by Moed (2020a), who cites the *Assessment of University Based Research* (AUBR) report of the European Commission (drafted by an Expert Group of which I have been member) as the main source (AUBR, 2010).

### *3.1. Unit of assessment*

By combining the provisions of the Law that created the Agency and the ministerial decrees that gave origin to the various VQR the following principles emerge.

#### 3.1.1. Institution-level assessment

The law that has regulated the Agency (DPR 1 febbraio 2010 n. 76) clearly states that the object of the activity is the quality of processes, results and products of activities of universities and Public Research Organisations (PROs), considering their internal configuration (art. 2, comma 1, lettera a). The units of assessment are clearly identified at institution level, i.e. universities and PROs.

The same text argues that in the assessment there must be consideration for disciplinary differences (art. 2 comma 3). This amounts to call for institutional level assessment with breakdown by discipline and/or internal administrative articulation. In practice, as described above, the evaluation is carried out at different institutional levels for universities and PROs. In universities, given that all researchers are by necessity classified by discipline and are affiliated to departments the assessment is carried out at (i) discipline; (ii) department; (iii) university-level. In PROs the assessment is carried out at institutional level and with a breakdown by internal organization (e.g. department or institute) but not by discipline, since researchers at PROs are not enrolled using the same classification system of universities (SSD).

#### 3.1.2. Prohibition of evaluation of individual researchers

It is important to remark that in all legislative and ministerial documents there is no mention at all of evaluation of *individual* researchers. This is the result of a hot political debate surrounding the law (although not documented in formal terms in the text), that clearly revealed a political orientation towards avoiding any possible use of indicators at the individual level.

This orientation is clear when compared to other provisions of the general legislative framework. For example, according to the university reform contained in Law 240/2010 universities can indeed evaluate their affiliates on an individual level, and use the results of assessment for monetary or other incentives, because they are the direct employers. Another provision that shows the prohibition to use the research assessment at the individual level is that the scores assigned to individual research products are communicated privately and electronically to their authors, while members of the Agency and experts are obliged to observe the strictest confidentiality. Finally, individual scores are not disclosed to anybody, following the general legislation on privacy.

#### 3.1.3. Involvement of all researchers

While researchers are not evaluated individually, with the VQR they must deliver the products to be evaluated to their institution (university and PRO) with an autonomous decision. This is in sharp opposition to the practice of VTR 2000–2003. This principle is evident from the provision that all in-service researchers (that is, all researchers having an employment contract with the university of PROs)[13] should be active, i.e. must submit a number of products. The practical implication is that the number of products to be evaluated per university is calculated by multiplying the number of in-duty researchers times the number of products per capita. The principle is also evident from the provision, in the ministerial decree for the first VQR, that researchers who did not submit products (so called *inattivi,* or non-active) contributed to the university score negatively, i.e. with a penalty.

Summing up, the Italian research assessment is clearly oriented towards the institution-level (university and PROs) but achieves this objective through the involvement of all researchers. The use of results of research assessment, as well as the use of metric indicators, for the evaluation of individual researchers is inhibited.

This principle results from the law that created the Agency and is evident in the parliamentary debate that addressed the issue before its approval. As a matter of fact, shortly after the first VQR some Rectors asked, formally or informally, to their professors to disclose voluntarily the individual score, or made some internal procedures conditional on individual scores. The Agency has systematically criticized this approach, stating explicitly that individuals cannot be evaluated with the research assessment methods used for large scale exercises. In the Report that presented the results of VQR 2004–2010 it was firmly stated that "the results of VQR cannot and should not be used to evaluate individual researchers". To the best of my knowledge these isolated mispractices disappeared, as an example of institutional learning.

### *3.2. Dimension of the research process*

#### 3.2.1. Definition of research performance

The overall orientation of the legislative and ministerial documents is towards a definition of research performance in terms of *scientific-scholarly impact*. The only term that is formulated and repeated across the texts is "quality". According

[13] This definition includes, for universities, Full professors, Associate professors, Researchers, both full time and part-time, and does not include post-doc researchers, doctoral students, and contract teachers. For PROs it includes researchers from all ranks and does not include technicians.

to Art. 3 (comma 1, lettera a) of the law (DPR 1 febbraio 2010 n. 76) the notion of quality is applied to "processes, results and products of activities of management, education and research".

In order to interpret this term appropriately we should make reference to other principles in the law. A prominent one is the reference to the best available practice at international level about the "evaluation of results" (Art. 1, comma 1). Art. 3 (comma 2, lettera b) states that the Agency evaluates "the quality of the products of research". It is clear that the focus of interest is not on the dimension of process (e.g. organization of research, management, resources, personnel, quality procedures), as is done for education in the context of quality assurance, but on the dimension of products, or results. The assessment or research is mainly an assessment of final results of an (unobserved) research process. These results must be materialized in specific objects, called research *products*. It must be remarked that, in the effort to interpret the policy mandate to cover all results, ANVUR defined the list of admissible research products in a very comprehensive way. In the first VQR basically imitated the long list of products of the UK RAE/REF. The list was then integrated with other items, particularly after suggestions of the experts in Architecture. Needless to say, the emphasis on research products implies that the model of research assessment is inevitably summative, not formative.

At the same time, other dimensions of the research process are mentioned in the law but with a non-systematic approach.

First, there is an explicit mention of technology transfer as a dimension of performance of research (Art. 3, comma 1, lettera a). As a matter of fact, the first ministerial decree launching the VQR included specific indicators of technology transfer. The decree (art. 6) asked institutions to provide data on:

(a) Patents and spinoffs of which the institution is owner or co-owner, with separate specification of the age and performance of spinoffs
(b) Cash inflows from sales and licensing of patents, with indication of the nature and characteristics of acquirers

It also dictated (art. 8) that the evaluation of patents should include transfer, development and socio-economic impact, including potential impact (Ministerial Decree 15 July 2011).

The decree did not add any other specification of indicators associated to third mission. ANVUR therefore added to the collection of institutional data an open section labelled Public engagement, inviting universities and PROs to give voluntary evidence of activities. The results were not used for the construction of indicators, given their lack of normalization.

As already stated, the role of social benefit or third mission of research has been suffering from a lack of formalization until the last VQR in 2019. The nature of evaluation of third mission in Italy and the methodological choices made for it will require a dedicated study.

Second, the final score attributed to universities includes not only the assessment of research products, but also the internationalization, the doctoral education and the quality of recruitment. In this way, several dimensions of the research process (somewhat similar to the concept of Infrastructure in the REF) are included. These indicators are evaluated at university or department level.

Summing up, there is a focus on the scientific-scholarly impact of research, with some consideration for the research process. The role of societal impact has been latent for some years but will be fully examined in 2021.

#### 3.2.2. Evaluation criteria

Specific evaluation criteria are not listed in the law, although it refers to the best international experience as a source. They are however clearly established in the ministerial decrees that have opened the VQR exercises. In this sense they satisfy the principle put forward by Moed (2020b) according to which "evaluation criteria and policy objectives cannot be founded informetrically". They have been defined in a policy-oriented document, put forward by the Ministry. As such, they work as constraints to the discretionary work of the Agency.

While they have been slightly changed across the three VQRs, they can be summarized as follows:

- Novelty or originality
- Relevance to the field, or impact
- Methodological rigor

The *originality* criterion (labeled "novelty" in the 2004–2010 edition) refers to a classical judgment of scientific authorship: the results of research must add knowledge to the state of the art. The criterion of *relevance* refers to scientific impact (i.e. relevance "to the field"), or the importance of the results of research for the overall scientific community. The judgment of the scientific community about the relevance of a particular piece of research speaks via the referees, or via the quantitative indicators. Again, please note that the notion of impact only refers to the scientific-scholarly dimension. The formulation has been changed into "attested or potential impact" in the 2011–2015 edition, in order to take into account some criticism from the humanities, in which the full impact of some piece of research may take many years to materialize and the internationalization of results may require lenghty processes of translation into foreign languages. In these cases the referees might reason in terms of potential, not realized impact. The formulation

"attested or potential", however, has been eliminated in the last edition, because it created some issue of interpretation in practice.

It is interesting to comment on the third criterion, *methodological rigor*. It substitutes the criterion *internationalization*, which was included in the first VQR 2004–2010 and eliminated subsequently. The meaning of internationalization as a quality criterion was to be interpreted in relative terms, that is, by defining the relevant scientific community in the largest possible sense in practice. The criterion, which was consistent with the overall orientation of the legislation towards the achievement of the best international standards, did not imply any devaluation of the research that is published in national language (as in humanities) or address a national community only (as in legal studies). Its application, however, was contested in some areas of SSH on the argument that a criterion labeled internationalization would implicitly (but strongly) lead referees to undervalue publications in Italian language. As I have demonstrated elsewhere, these arguments do not have a strong empirical support. While in general articles in English received on average better scores than articles in other languages (Ferrara et al. 2016), it is also true that research results in Italian language and in book format in those areas in humanities in which these results are considered state of the art received top scores (Bonaccorsi, 2016). As a matter of fact, this criterion was substituted by a new one, labelled methodological rigor, for which the interpretation is less controversial. The overall discussion about this criterion is described at lenght in Bonaccorsi (2018).

#### 3.2.3. Evaluation methodologies

While the law regulating the Agency (DPR 1 febbraio 2010 n. 76) is silent on specific evaluation criteria and delegates them to ministerial decrees, it is prescriptive on issues of methodology. This rather surprising aspect of the evaluative framework can be explained with reference to the opening part of the law, which is part of the interpretive apparatus of the law itself. In this section the law mentions all other laws that must be considered in order to interpret the text and the positions of the legislative bodies that have had a role in the approval. In one of these positions the members of Parliament (House of Representatives) of Commission VII (Culture and education) asked to modify the text of art. 3 (comma 2 lettera b) according to which the evaluation of research products is carried out "mainly (*principalmente*) with procedures of peer review". Their suggestion was to eliminate the adverb, so that peer review would be qualified as the unique methodology.

Interestingly, the law rejects this position, arguing that there might be sectors and cases in which evaluative informetrics is admissible ("è ammissibile la valutazione metrica"). Assuming the peer review as the only admissible methodology, according to the law, would have limited the discretionary appreciation of the Agency in the selection of the best methodology. The law therefore establishes that the Agency "utilizes the criteria, the methods and the indicators which are considered more appropriate for any type of evaluation" (art. 3 comma 3).

Again, I find that this provision is consistent with the requirement that the two levels of research assessment (Policy and management, and Evaluation) should define the objectives, purposes and criteria but not the methodologies, while the research assessment exercise (Analytics and Data collection) should be responsible for the methodological and technical choices.

It is interesting to place this legislative text in the context of the large debate on peer review vs informetrics in research assessment, ignited by *The metric tide* report (Wilsdon et al. 2015). My reading of the debate is that it has been channeled too narrowly into the issue of whether informetrics might replace peer review. The fierce reaction against metrics has been magnified by the need to resist against an assumed final attack of bibliometrics against the principles of peer review. As a matter of fact, some suggestion for the exclusive use of peer review has been advanced in the international literature (Harzing, 2018). This line of argumentation clearly influenced the position of Commission VII of the Italian Parliament, who tried to prohibit bibliometrics by law. As a matter of fact, the idea that bibliometrics is dangerous, while peer review is virtuous is largely held: there is currently a large literature on the disadvantages of metrics, while the disadvantages of peer review are studied much less.

The Italian law took a balanced position. Both methodologies are admissible (peer review and metrics), so that any position forbidding one of the two is not acceptable. Peer review must prevail in aggregate terms, however. At the same time, they must be adopted with specific reference to the needs of disciplines. And the final choice must be left to the independent Agency, not the government or the Parliament.

### *3.3. Purposes and objectives*

It is possible to derive a number of principles and statements about the objectives and purposes of research assessment.

#### 3.3.1. Accountability

The law creating the Agency (Legge 24 novembre 2006, n. 286) and the law regulating its activities (DPR 10 febbraio 2010, n. 276) are clear in stating that research assessment should be external, impartial, independent, professional, transparent and public. It must be directed towards the evaluation of effectiveness and efficiency of all research activities funded with public money. These formulations clearly point to a major objective of accountability. Institutions that receive public money must be able to justify their results.

This formulation, which is in some sense a standard tenet of public policy towards accountability, in the context of Italian policy making had a peculiar meaning. It came after several years of debate and policy efforts to reform the Public Administration and was introduced into one of the sectors of the administration in which accountability was traditionally very low, given the power of academic staff. It was then associated to large expectations from the public opinion and the media.

#### 3.3.2. Performance-based funding

The two mentioned laws are also explicit in mentioning the use of research assessment results as an input for the allocation of funding to universities. The law 24 novembre 2006, n. 286 states that "the results of the evaluation activities of ANVUR constitute the reference criteria for the allocation of public funds to universities and research institutes" (comma 139). This statement is repeated in DPR n. 276 (art. 4 comma 1).

As a matter of fact, the results of evaluation are used only for a small part of the overall funding, the so called *quota premiale*. In addition, in the final allocation the funding formula adopted by the Ministry has been associated to a clause, according to which the loss in funding for the under-performing universities is kept under a given threshold. As a result the *average* impact of performance-based funding is much smaller than it is in other countries, notably in UK.

At the same time, the *marginal* impact has been important. It has raised drastically the level of awareness and attention of university administrators about the importance of performance in research quality.

In addition, the provision that the results of research assessment must enter into the allocation of funding has the consequence that they must be formulated in a quantitative way. Consequently, the ministerial decrees that have launched the various VQRs have defined a quantitative framework and the Agency has published results accordingly.

#### 3.3.3. External communication

As a corollary of transparency and publicness of the overall evaluation there is also an explicit objective of improved communication with respect to stakeholders, in particular students.

### *3.4. Relevant, general or "systemic" characteristics of the units*

This dimension of the evaluative framework is difficult to examine using the official documents. I can mention a systemic issue that is peculiar to the Italian economic and social background, i.e. the North-South divide. As we will see below and in the companion paper, there are large differences across universities located in the Northern and Southern regions of the country. Given that the research assessment has implications on funding, this issue has raised large attention.

My reading of the legislation is that this problem is not addressed in the context of the traditional notion of *cohesion*, as it happens for Structural Funds of the European Commission and for other policies. The word cohesion does not appear in legislative texts. The implicit assumption is that universities that operate within the same public funding framework should be treated equally in terms of accountability. For a longer discussion of this issue see the companion paper.

## 4. The impact of VQR on universities. Learning how to deal with rankings

The first VQR was certainly a shock for the academic system. Not only for the first time all researchers were asked to submit their products, but the final results showed several unexpected facts with respect to the ranking of universities. Moreover, the financial impact of the evaluation was significant and academic bodies and university administrators started to be worried about the implications for their annual budget.

In particular, Law 98/2013 stated that the performance-based funding of universities should be an increasing share of the overall funding (FFO, *Fondo di Finanziamento Ordinario*), starting from 16% in 2014, 18% in 2015 and 20% in 2016, with a maximum share of 30% to be achieved gradually. The share of performance-based funding should be based on VQR results for 3/5 of the total. The same law stated that the maximum decrease of funding for each university, should be minus 5% with respect to the previous year. To understand better the impact on the funding of universities, it must be reminded that government core funding is still approximately 80% of total revenues (EUA 2018; ETER, 2019). For many universities the (almost) fixed expenditure for personnel absorbs the large majority of funding. This means that the *marginal* impact of a reduction in government funding is significant.

In subsequent years the impact of research assessment percolated down all layers of the academic world. From this perspective the impact of VQR on university administrators and the overall academic body has been deep and pervasive.

In 2013 universities were unprepared to manage the sudden visibility given by the VQR. When the results were made public, the media started a campaign which lasted several weeks and involved hundreds of headings, with a coverage which was much larger than for any academic topic. In a couple of papers we have examined the impact of VQR on universities assuming the perspective of coverage of the overall exercise and of ranking of individual institutions in the media. Blasi et al. (2017) showed that the media selected only the information which was framed in terms of ranking and focused on a restricted portion of the rank, that is, the first three positions, or "the best universities", with a framing similar to Olympic games. Non-ranking information was filtered out. Universities located below the podium were rarely mentioned. On the contrary, large coverage was given to the bottom three of the ranking, or "the worst universities". It

was clear that the selection of content was done by the media, while universities were mostly passive. Blasi et al. (2018a) further examined the media coverage and found that the number of occurrences of individual universities in articles was only explained by the Olympic factor, or the number of times they appear in the podium, or in any of the top three positions. The media used the VQR to build up a game representation, dominated by a winner-takes-all logic, as it happens for celebrities.

Interestingly, the situation changed significantly after the second VQR. When the data were presented in 2017, universities had enough experience to anticipate the main results. In most cases they had invested into communication offices that were prepared to exploit the news in a self-interested way. Bonaccorsi et al. (2020a; 2020b) replicated the same regression models used in Blasi et al. (2018a) with the media coverage in 2017 and found a significant reduction in the celebrities effect, or the disproportionate attention given to the top universities. In particular, they found that universities systematically adopted categorization tactics: they self-defined a collection of partial rankings in which their position was at the top. These self-made rankings were defined in geographic terms (by region, by macroarea of the country), by institutional nature (public vs private), by discipline (generalist vs specialist), by size (small, medium or large). By combining various dimensions it was possible to design partial rankings and to launch a communication campaign in which the good position of the university could be announced and made relevant. In this way the media received press releases, which were transformed into articles, from universities that were not positioned at the top (that is, they were not candidates to receive attention in the global ranking) but could claim a top position in one of the many self-made partial rankings. This communication was clearly aimed at confirming the loyalty of stakeholders, including students and their families.

There is a large international debate about the impact of university rankings and the risks associated to the creation of vertically stratified higher education systems. The intriguing findings about categorization tactics by Italian universities show that counterbalance moves are actively implemented. Universities try to prevent the creation of a higher education system which is perceived as stratified in terms of research excellence by focusing on other dimensions of performance and communicating systematically with their stakeholders. In other words, those universities that know they cannot compete on research excellence try to differentiate their offering. It is probably too early to evaluate whether this differentiation increased after the introduction of research assessment.

## 5. The new recruitment system and the median saga

After the legislative initiative for the creation of the Agency in 2010 the Parliament approved a major reform of the academic recruitment system. The history of various solutions experienced in the Italian system is quite complex and cannot be outlined here.

It is however useful to reconstruct briefly the cultural climate that led to the 2010 reform and its implementation. Complaints about the lack of transparency and meritocracy of academic promotions have been repeatedly made in the scientific literature (Martin, 2009) and have been even the object of parliamentary initiatives (Fréville, 2001). These issues have been often reported in the case of Italy (Perotti, 2008; Allesina, 2011; Ferlazzo and Sdoia, 2012; Pezzoni, Sterzi and Lissoni, 2012; Abramo, D'Angelo and Rosati, 2014; Bagues, Sylos Labini and Zinovyeva, 2015). In the last thirty years the news about favoritism in Italian academic recruitment have attracted the attention of *Nature* (Gaetani and Ferraris, 1991; Amadori et al. 1992; Aiuti, Bruni and Leopardi, 1994), *Science* (Biggin,1994), and *Lancet* (Fabbri, 1987; Gerosa, 2001; Garattini, 2001), among others. Episodes of nepotism and familism in academic recruitment, when reported in the press, generate a wave of resentment in the public opinion, as if the academy were a privileged caste, using public money in their personal interest (Durante, Labartino and Perotti, 2011).

This background helps to understand the origins of the legislation that went into service in 2010, introducing a radical departure from the tradition (Capano and Rebora, 2012). As a matter of fact, the Law 240/2010 introduced a dual-layer recruitment system. At university level the final decision to recruit academic staff is made, after a proposal of departments and the approval and financial authorization of the Board (*Consiglio di Amministrazione*). The new system, made operational in 2012 after a Ministerial decree and still in place (with several modifications[14]), is based on the combination of a National Scientific Habilitation at a national level and a local or university-based system of recruitment or promotion. The National Habilitation, following the French and Spanish experiences, is managed at a national level by the Ministry, in order to ensure a uniform level of *minimal* requirements for the academic career. Among those receiving the Habilitation, and only among them, universities could hire candidates. The requirement for applying to the Habilitation procedure is that candidates overcome a quantitative threshold in indicators of academic performance. After qualifying according to these indicators, candidates will be evaluated qualitatively by the committee, taking into account the entire scientific production, as well as a list of academic and institutional achievements. Marzolla (2015) offers a very detailed quantitative reconstruction of the procedure.

There were several new provisions. The committee was formed by five full professors, but one of them should not be affiliated to Italian universities, but to a university or equivalent institution in OECD countries. This provision was intended to increase the transparency and to mitigate the supposedly deep propensity of Italian academicians to make

[14] Decreto del Presidente della Repubblica 14 settebre 2011, n.222. The law 240/2010 was subsequently modified by the Law 114/2014. The original decree, which went into service in 2012, was deeply modified in 2016 by the Decreto del Presidente della Repubblica 4 aprile 2016, n. 95.

collusive agreements within the committees.[15] Committees should publish their quality criteria before examining the list of candidates, and candidates were permitted to withdraw their application after the publication of criteria. All candidates made their CV available on the official website of the procedure, together with the list of publications. In addition, the legislation established a long list of detailed quality criteria to be met by all candidates, irrespective of the discipline. It is important to note that the decision of granting the Habilitation required four positive votes out of five members. The selection of high quality members of the Habilitation committee, coupled with published criteria and a high threshold for positive votes was intended to raise the overall transparency of the procedure.

In fact, the most radical innovation was that not all Full professors were admitted to the competition for becoming members of the Habilitation committee, but only a selection of them- as a matter of fact, only 50% of them, as we shall see. For the first time in the history of the academic community in Italy, not all Full professors were considered legitimate as members of committees for the selection and cooptation of colleagues. Full professors submitted their application for being included in the list of candidates. The committees were formed by random sampling within the lists of candidates. To be included in the list, however, applicants should satisfy a number of requisites established in the decree. The general principle was that members of the Habilitation committees should be at least of equal value of the potential candidates. There were two types of requisites: one based on the qualitative appreciation of a number of activities (recipient of research grant, project coordination, editorial work and the like), the other based on quantitative indicators. The indicators were established in a detailed way in the decree. They were divided in two groups: bibliometric and non bibliometric indicators. The former were applied to STEM disciplines, the latter to SSH. All indicators were to be computed for the entire scientific career of the candidates.

Bibliometric indicators were as follows: (1) number of articles in indexed journals; (2) number of citations received; (3) h-index. Non bibliometric indicators were instead: (1) number of books; (2) number of book chapters and articles; (3) number of articles in A-rated journals. All indicators were normalized by academic age, defined as the number of years since the first publication. Given the quantitative nature of these indicators, there was a need to specify the threshold value. The decree established that the threshold was to be computed as the median value of the distribution of the indicators across the entire academic community. In particular, candidates to the Associate professorship should meet at least the median value of the indicators for those already affiliated as Associate professors; candidates to Full professorship should meet the median value of those already in this position. In bibliometric sectors candidates had to overcome at least two out of three indicators, while in non bibliometric sectors the requirement was only one.

To clarify the meaning of the median, researchers who wanted to be recognized as Associate professors in STEM disciplines should have more citations than 50% of the current Associate professors in the same discipline. The median has robust statistical properties, insofar as it is not influenced by extreme values. Establishing the median value as the threshold had two dramatic consequences: first, it was a severe requirement; second, it was dynamically adjusted upwards, since newly admitted professors would have by definition better indicators than the existing population. ANVUR strongly defended in official documents the use of the median value as a device for inducing qualified recruitment.

The median value of these indicators was to be computed by ANVUR. The Agency published detailed criteria[16] for the computation of indicators: for example, indexed journals were to be taken from Web of Science or Scopus. In STEM disciplines the calculation of median values was made easier by the availability of WoS and Scopus (candidates were entitled with the more favourable count). It turned out that normalization by scientific disciplines as defined in administrative terms (rather than by Subject Category of journals) might produce distortions. For example, most professors in Physics share the same large SSD (*Settore Scientifico Disciplinare*), but the median value of all bibliometric indicators is completely different depending on whether they work in Particle Physics or Matter Physics, given the large differences in experiments and co-authorships. Or professors in Computer Science largely differ in terms of their theoretical or applied orientation. Taking into account observed differences in the distribution of indicators, ANVUR published single median values for most disciplines, and multiple median values in a few cases.

More complex was the production of indicators in SSH. To start with, there was no definition of a scientific publication. The list of journals was extracted from a dataset, managed by the University Consortium CINECA, originated from loginmiur, or the self-managed personal website of all academic staff affiliated to universities, in which people record their own publications. Preliminary analysis, in fact, showed that researchers used to fill metadata with all sorts of publications, from newspaper articles to local bulletins, from grey literature to narrative books. No definition of the eligible journals was included in the dataset. Metadata are not curated according to library criteria. In Summer 2012 CINECA collaborated with ANVUR in extracting all metadata and examining their distributions.

For non-indexed journals, ANVUR was in charge of classifying *all* journals in which Italian scholars had published in the 2002–2012 period. The provision to cover all journals "in which Italian scholars have published" was intended to be comprehensive with respect to the scientific production of SSH disciplines using national language and publishing in national journals, or more generally in non-indexed journals. In practice, the classification was a daunting exercise:

[15] This provision was eliminated in 2016 due to various administrative problems (e.g. lack of comparability across disciplines between national and foreign academics).

[16] All documentation available at https://www.anvur.it/attivita/asn/.

according to a document published by ANVUR,[17] as many as 42,494 titles were examined, corresponding to 15,998 journals. Of these, 12,865 were classified as scientific journals, eliminating 3,133 non scientific journals (19.6%).[18]

On top of this, the Agency was asked to rate scientific journals in order to obtain a category, called A-rated journals, on which the third indicator was based. The total number of journals initially rated A-class was 3,676. This indicator (Number of articles in A-rated journals) was in fact the only one in non-bibliometric sectors that was explicitly based on quality of research criteria, while the two others (Number of books; Number of book chapters and journal articles) were mainly volume-based. Remember that there was no restriction on the type of books (hence, of chapters of books) to be included in the indicator.

For the sake of the current discussion, it must be remembered that the classification was carried out in two months only (*sic*), upon mandate of the Ministerial decree. ANVUR nominated an expert group of 28 scholars, with a mandate to solicit and take note of the opinion of learned societies. Inevitably, there have been tensions between the Agency and scientific communities, which were involved through their respective learned societies in the production of informed opinions on the list of scientific journals and of A-rated journals, respectively. This exercise, however, was done under time pressure, given the huge expectations of the academic community for initiating the Habilitation procedures and starting the recruitment of new staff, after many years of hiring freeze. Inevitably, the initial classification, which was published within the deadline, included a few dozens misclassification errors. Even if these errors represented just 0,12% of journals (one out of one thousand) and were rapidly corrected they gave origin to a hostile media campaign, which achieved the headings of a national newspaper.

The lists of indicators were published between the end of August and the start of September 2012, following the deadline of the Ministerial decree. They were published separately for candidates to the Habilitation, based only on the decade 2002–2012, separately for Associate and Full professorships, and for candidates to the Habilitation committees, based on the entire scientific career of Full professors. For each of these three categories, the lists included three bibliometric and non-bibliometric indicators, represented by the median value of the distributions. Full professors applying to a Habilitation committee in bibliometric sectors should meet at least two of the three indicators (i.e. overcome the median value of their peers) while in non bibliometric sectors the requirement was only for one indicator.

After the publication of indicators Full professors submitted their candidature as members of the Habilitation committees. They submitted their CV and the full list of publications. Based on these data, the Agency informed them about the admission to the final list for the random extraction. In order to facilitate the procedure, the Agency had confidentially informed all Full professors of their position with respect to indicators. All Full professors received in their own personal website a message stating whether their list of publications included publications that resulted above or below the three median values. This information was provided in a traffic light form, where "green" meant that the indicator was satisfied, and "red" that it was not.

The traffic light notwithstanding, several Full professors submitted their candidature irrespective of the indicators. The number of applications was 7325, a very large number. Of these, 1468, or 20% were rejected. The Agency sent a personal mail to all candidates, stating whether each of the indicators were satisfied or not, and concluding on whether the candidature was eligible or not (that is, whether at least two indicators for bibliometric sectors and one for non-bibliometric sectors were satisfied).

In order to take into account possible mistakes and to manage controversies, rejected candidates could submit an appeal. After controlling manually for these cases, a few exclusions were corrected. Summing up, 20% of full professors were told they were not in the position to serve as members of the committees for the cooptation of other colleagues in the academic community. This was a shocking novelty. It had never been done before in the history of Italian universities. An external authority, the Agency, was made responsible for entering into the sacred domain of academic autonomy and self-reference in order to admit or reject professors, who were at the acme of their scientific career.

It is useful to examine the composition of the 1468 candidatures that were rejected. The largest group comes from life sciences: 305 in Medicine, 151 in Biology, 137 in Veterinary sciences. Almost 600 Full professors in these disciplines were gently told they were not legitimated to sit in the Habilitation committee. Next comes Engineering, with 163 rejected candidatures. Third comes Arts and Humanities with 104, then Law with 97 and Architecture with 95. Less represented in the top list are hard sciences.

By and large, rejected candidates were those with a poor scientific production, as compared with national colleagues of the same discipline. In some cases, however, they were scientifically active, but published in non-eligible sources. For example, medical researchers publishing books or articles in Italian journals suddenly found their production was not eligible, because in bibliometric sectors only indexed journals were relevant. A similar pattern was found in traditional areas of engineering, in which textbooks are considered scholarly products, but were deliberately ignored in the calculation of indicators. It is also interesting that a large proportion of rejected professors come from applied disciplines, in which academicians are also involved into professional and consulting activities, as it happens in Medicine,

[17] Consiglio Direttivo ANVUR, *Chiarimenti sulla classificazione delle riviste nell'ambito della abilitazione scientifica.* 5 ottobre 2012.

[18] As a matter of fact, the same journals appeared repeatedly across many scientific areas. While the judgment of lack of scientific nature had to be valid universally (although there were cases of controversies across experts for journals of high culture), the A-class rating was clearly specific to the disciplines and might lead to different decisions. This means that the same journal was subject to many evaluations. My own reconstruction from internal data is that the total number of items evaluated was 67,038 titles.

Engineering, Law or Architecture. Broadly speaking, we may expect that rejected candidates were: (a) no longer active researchers; (b) mostly traditional scholars, publishing in non eligible sources; (c) less productive researchers, either due to a heavy engagement into external, non academic activities, or due to intrinsic lower productivity. In all three cases the reasons for exclusion seemed well grounded in the legislative mandate, which explicitly aimed at establishing international standards and transparent procedures in recruitment.

While the rationale for this provision is clear, nevertheless the technique used was a radical departure from the tradition in the history of the relations between the State and academic communities (Marini, 2017). In the 20th century, particularly after the experience of totalitarian regimes, the notion of autonomy of scientific research has been developed in almost all European countries in their constitutional laws, or at least in primary legal sources. And indeed there has been a legal opinion, supported by the learned society in Constitutional Law, that has argued against the possibility to restrict the admission of Full professors to committees. An appeal to the Administrative Court, however, was rejected: the Court made clear that the decree was perfectly legitimate in defining the requirements for becoming members of the Habilitation committees, provided that these requirements were transparent and controllable. In addition, since the Agency was established by means of a law, it was considered that its authority was fully enforced in the legal system and demanded compliance by all interested parties.

The procedures were modified in 2016[19] and the new procedures have been adopted for the 2016–2018 and 2018–2020 rounds of National Habilitation.

The most important changes are as follows:

- The Habilitation procedure has been made permanent: Habilitation commitees are formed each two years but candidates may submit their candidature at any time, and the Habilitation has validity six years;
- the committee is formed by five full professors (no foreign member);
- the majority requested is three votes out of five;
- it is possible to establish the maximum number of publications to be submitted for evaluation;
- the criteria and quantitative parameters should be defined with a Ministerial Decree, after proposal from ANVUR and CUN (*Consiglio Universitario Nazionale*, or National University Council).

The most important changes refer to the majority vote (three members of the committee instead of four) and the decision making process for the criteria and parameters. In the 2012 application of the law, the responsibility for the definition of criteria and parameters was entirely delegated to ANVUR, a technical independent body. In the 2016 reform the procedure for the publication of indicators changed. Instead of being delegated to ANVUR, as it was in 2012, the numerical value of indicators has been decided by the Ministry after consultation with ANVUR, but also with an elective body (CUN), in which all categories of academic staff (researchers, associate and full professors, as well as students) are represented.

As a matter of fact the median values were eliminated. They were substituted by thresholds, or quantitative values for each of the indicators that represented minimum requirements. The statistical foundations of the thresholds have not been published. They are based, as it was in 2012, on the distribution of indicators across the population of university researchers, as maintained by CINECA. At the same time there is no way to identify a statistical concept that has been applied in a uniform way.

The Ministry published tables with the thresholds in 2016 and 2018. By inspecting the tables and comparing them with the median values published in 2012 it appears that the thresholds are much lower, perhaps placed at the 30% of the distribution. It is too early to compare the effect of the 2012–2013 procedures and the 2016–2020 ones. I have obvious conflict of interest in arguing in favour of the median approach but I must observe that the wave of criticism against the Habilitation settled down after 2016. The procedures are running smoothly. In the companion paper I will discuss some recent studies that have evaluated the ASN procedure.

## 6. Policy highlights

From the Italian experience I think that a number of policy highlights can be derived.

First, it is important that research assessment is prepared by an extensive period of discussion with scientific communities and institutions. This discussion should cover not only the broad policy goals, but also research assessment criteria and methodologies. Most likely the academia will not have developed detailed understanding of the methodological and technical issues at stake. What is required is not deliberation, but involvement. In the Italian experience this open discussion was made almost impossible by the extremely long delay of implementation of the legislative provision (from 2006 to 2011) and the need to bootstrap the implementation. As a general recommendation, however, an adequate preparation period is to be planned.

Second, if research assessment takes place in a period of financial restrictions to research and/or higher education, it is perceived as a policy instrument to reduce resources by avoiding to take the political responsibility. In other

[19] Decreto del Presidente della Repubblica 4 aprile 2016, n. 95; Decreto Ministeriale 7 giugno 2016, n. 120; Decreto Ministeriale 29 luglio 2016, n. 602.

experiences of European countries, namely United Kingdom and the Netherlands, the introduction of research assessment has been associated to an increase of financial resources in real terms. Performance-based funding is then perceived as an incentive to perform better. Again, the Italian experience shows the negative impact of implementing research assessment in a period of budget crisis. The share of funding based on research performance is not perceived as a prize, but as a reduction of penalty.

Third, linking the research assessment to the allocation of resources ensures a fast and pervasive impact. In a few years academic leaders and administrators have learnt to take seriously into account the results of the assessment. This awareness percolates rapidly to the academic community. There is an international debate on the pros and cons of performance-based funding in which the Italian experience is not yet examined, due to the short time window of the implementation. However, the comparison between an exercise without financial implications (VTR 200–2003) and one with implications on the funding formula (VQR 2004–2010) clearly shows the superiority of the latter.

Fourth, quantitative indicators help a lot. They work as focusing devices and approximate the underlying complex constructs (such as research quality) in such a way to generate attention and awareness. We must be aware, at the same time, that a number of researchers simply do not have the training and attitude to reason quantitatively in a smooth way.[20] They perceive quantitative indicators as an external imposition, whose meaning is uncertain. There is a need for an extended work of clarification and practice. This is particularly true in administrative and political cultures, such as the Italian one, in which the attitude to make decisions based on numbers is historically weak.

Fifth, policy makers and research assessors should be prepared to manage conflicts. Although the normative foundations for research assessment may actively embrace ideals of neutrality and professionalism, there is always room for conflicts. They come from a variety of sources, some of which open (e.g. conflicts on assessment criteria, or political commitment against assessment), some covert (e.g. protection of vested academic interest). The opposition against research assessment merges together these sources. This makes it extremely difficult to interpret the intentionality of opponents. My recommendation is to put in place systematic and transparent procedures of hearings and discussions. The arguments of opponents should be spelled out and rationally debated. Conflicts must be managed with open listening and open assumption of responsibility.

Finally, policy makers should be keenly aware of the unintended consequences of research assessment and prepare counterbalance actions. Perhaps the most risky consequence is bureaucratization. The slow transformation of research assessment into a routine administrative activity is dangerous. Another unintended consequence is the adoption of strategic or gaming behaviors to a point where the academic values are placed at risk. Research assessment policies and practices should be periodically re-opened and re-motivated in order to prevent or mitigate these outcomes.

## Competing Interests

The author has no competing interests to declare.

## References

**Abramo, G., D'Angelo, C. A.,** & **Caprasecca, A.** (2009). Allocative efficiency in public research funding: Can bibliometrics help? *Research Policy*, *38*(1), 206–215. DOI: https://doi.org/10.1016/j.respol.2008.11.001

**Abramo, G., D'Angelo, C. A.,** & **Rosati, F.** (2014). Career advancement and scientific performance in universities. *Scientometrics*, *98*(3), 891–907. DOI: https://doi.org/10.1007/s11192-013-1075-8

**Aiuti, F., Bruni, R.,** & **Leopardi, R.** (1994). Impediments of Italian science. *Nature*, *367*(6464), 590. DOI: https://doi.org/10.1038/367590a0

**Allesina, S.** (2011). Measuring nepotism through shared last names: The case of Italian academia. *PLoS ONE*, *6*(8), e21160. DOI: https://doi.org/10.1371/journal.pone.0021160

**Amadori, S., Bernasconi, C., Boccadoro, M., Glustolisi, R.,** & **Gobbi, M.** (1992). Academic promotion in Italy. *Nature*, *355*(6361), 581. DOI: https://doi.org/10.1038/355581a0

**Ancaiani, A., Anfossi, A. F., Barbara, A., Benedetto, S., Blasi, B., Carletti, V., Cicero, T., Ciol, A., Costa, F., Colizza, G., Costantini, M., di Cristina, F., Ferrara, A., Lacatena, R. M., Malgarini, M., Mazzotta, I., Nappi, C. A., Romagnosi, S.,** & **Sileoni, S.** (2015). Evaluating scientific research in Italy: The 2004-10 research evaluation exercise. *Research Evaluation*, *24*, 242–255. DOI: https://doi.org/10.1093/reseval/rvv008

**Anfossi, A., Ciolfi, A., Costa, F., Parisi, G.,** & **Benedetto, S.** (2016). Large-scale assessment of research outputs through a weighted combination of bibliometric indicators. *Scientometrics*, *107*, 671–683. DOI: https://doi.org/10.1007/s11192-016-1882-9

**AUBR.** (2010). Assessment of University-Based Research Expert Group (AUBR). Assessing Europe's University-Based Research. *K1-NA-24187-EN-N, European Commission*, Brussels. http://ec.europa.eu/research/era/docs/en/areas-of-actions-universities-assessing-europeuniversity-based-research-2010-en.pdf

[20] This was clear after the introduction of the notion of "median" in the National Scientific Habilitation. It took the academic community by surprise. I have an entire collection of anecdotes of respected authorities that made declarations in which the difference between the median and the mean of a distribution was ignored. They simply did not know it.

**Biggin, S.** (1994). Corruption scandal reaches academe. *Science*, *266*(5187), 965. DOI: https://doi.org/10.1126/science.266.5187.965

**Blasi, B., Bonaccorsi, A., Nappi, C.,** & **Romagnosi, S.** (2019). The link between research quality and technology transfer in the Italian Evaluation of Research Quality VQR 2011–2014. *Paper presented to the ISSI Conference*, Rome, 2–5 September.

**Blasi, B., Romagnosi, S.,** & **Bonaccorsi, A.** (2017). Playing the ranking game: media coverage of the evaluation of the quality of research in Italy. *Higher Education*, *73*, 741–757. DOI: https://doi.org/10.1007/s10734-016-9991-1

**Blasi, B., Romagnosi, S.,** & **Bonaccorsi, A.** (2018a). Universities as celebrities? How the media select information from a large Research Assessment Exercise. *Science and Public Policy*, *45*(4), 503–514. DOI: https://doi.org/10.1093/scipol/scx078

**Blasi, B., Romagnosi, S.,** & **Bonaccorsi, A.** (2018b). Do SSH researchers have a third mission (and should they have)? In A. Bonaccorsi (ed.), *The evaluation of research in Social Sciences and Humanities. Lessons from the Italian experience* (pp. 361–392). Dordrecht: Springer. DOI: https://doi.org/10.1007/978-3-319-68554-0_16

**Bonaccorsi, A.** (2016). L'impatto della valutazione sulle scienze sociali in Italia. Lo strano caso delle discipline aziendali e della sociologia. *Notizie di Politeia*, *n. 123*, 36–45.

**Bonaccorsi, A.** (2018). Peer review in Social Sciences and Humanities. Addressing the interpretation of quality criteria. In A. Bonaccorsi (Ed.), *The evaluation of research in Social Sciences and Humanities. Lessons from the Italian experience* (pp. 71–102). Dordrecht: Springer. DOI: https://doi.org/10.1007/978-3-319-68554-0_4

**Bonaccorsi, A.** (Ed.). (2018). *The evaluation of research in social sciences and humanities*. Dordrecth: Springer. DOI: https://doi.org/10.1007/978-3-319-68554-0

**Bonaccorsi, A., Belingheri, P., Blasi, B.,** & **Romagnosi, S.** (2020a). Institutional responses to university rankings. A tale of adaptation and cognitive framing. Forthcoming In E. Hazelkorn (Ed.), *Research handbook on university rankings: History, methodology, influence and impact.* Cheltenham: Edward Elgar.

**Bonaccorsi, A., Belingheri, P., Blasi, B.,** & **Romagnosi, S.** (2020b). Self-made university rankings. Categorization tactics and communication activism in Italian universities. Forthcoming, Research Evaluation.

**Capano, G.,** & **Rebora, G.** (2012). Italy: From bureaucratic legacy to reform of the profession. In G. Philip, A. L. Reisberg, M. Yudkevich, G. Androushchak & I. F. Pacheco (Eds.), *Paying the professoriate. A global comparison of compensation and contracts*. New York: Routledge.

**Cuccurullo, F.** (2006). La valutazione triennale della ricerca VTR del CIVR. Bilancio di un'esperienza. *Analysis. Rivista di cultura e politica scientifica* (pp. 3–4, 5–7).

**Durante, R., Labartino, G.,** & **Perotti, R.** (2011). Academic dynasties: Decentralization and familism in the Italian academia. *NBER working paper series* (pp. 17572). DOI: https://doi.org/10.3386/w17572

**ETER.** (2019). How are European Higher Education Institutions funded? New evidence from ETER microdata. *ETER Analytical Report 02*. Available at https://www.eter-project.com/#/analytical-reports.

**European University Association.** (2018). Public Funding Observatory Report 2018, Brussels.

**Fabbri, L. M.** (1987). Rank injustice and academic promotion. *Lancet*, *2*(8563), 860. DOI: https://doi.org/10.1016/S0140-6736(87)91051-8

**Ferlazzo, F.,** & **Sdoia, S.** (2012). Measuring nepotism through shared last names: Are we really moving from opinions to facts? *PLoS ONE*, *7*(8), e43574. DOI: https://doi.org/10.1371/journal.pone.0043574

**Ferrara, A.,** & **Bonaccorsi, A.** (2016). How robust is journal rating in Humanities and Social Sciences? Evidence from a large-scale, multi-method exercise. *Research Evaluation*, February 2016. DOI: https://doi.org/10.1093/reseval/rvv048

**Franceschet, M.,** & **Costantini, A.** (2011). The first Italian research assessment exercise: A bibliometric perspective. *Journal of Informetrics*, *5*, 275–291. DOI: https://doi.org/10.1016/j.joi.2010.12.002

**Fréville, Y.** (2001). *La politique de recrutement et la gestion des universitaires et des chercheurs*. Rapport d'information n°54, Sénat, 6 Novembre 2001.

**Gaetani, G. F.,** & **Ferraris, A. M.** (1991). Academic promotion in Italy. *Nature*, *353*(6339), 10. DOI: https://doi.org/10.1038/353010a0

**Garattini, S.** (2001). Competition for academic promotion in Italy. A reply. *Lancet*, *357*(9263), 1208. DOI: https://doi.org/10.1016/S0140-6736(00)04357-9

**Gerosa, M.** (2001). Competition for academic promotion in Italy. *Lancet*, *357*(9263), 1208. DOI: https://doi.org/10.1016/S0140-6736(00)04356-7

**Harzing, A. W.** (2018). Running the REF on a rainy Sunday afternoon: Can we exchange peer review for metrics? Leiden: STI 2018 Conference Proceedings (pp. 339–345).

**Marini, G.** (2017). New promotion patterns in Italian universities: Less seniority and more productivity? Data from ASN. *Higher Education*. DOI: https://doi.org/10.1007/s10734-016-0018-8

**Martin, B.** (2009). Academic patronage. *International Journal for Educational Integrity*, *5*(1), 3–19. DOI: https://doi.org/10.21913/IJEI.v5i1.478

**Marzolla, M.** (2015). Quantitative analysis of the Italian National Scientific Qualification. *Journal of Informetrics*, *9*, 285–316. DOI: https://doi.org/10.1016/j.joi.2015.02.006

**Perotti, R.** (2008). *L'università truccata.* Torino: Einaudi.

**Pezzoni, M., Sterzi, V.,** & **Lissoni, F.** (2012). Career progress in centralized academic systems: an analysis of French and Italian physicists. *Research Policy*, *41*(4). DOI: https://doi.org/10.1016/j.respol.2011.12.009

**Wilsdon, J., Allen, L., Belore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Kain, R., Kerridge, S., Thelwall, M., Tinkler, J., Viney, I., Wouters, P., Hill, J.,** & **Johnson, B.** (2015) *Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management.* London: Higher Education Funding Council for England. DOI: https://doi.org/10.4135/9781473978782

**How to cite this article:** Bonaccorsi, A. (2020). Two Decades of Experience in Research Assessment in Italy. *Scholarly Assessment Reports*, 2(1): 16. DOI: https://doi.org/10.29024/sar.27

**Submitted:** 15 September 2020 **Accepted:** 29 October 2020 **Published:** 17 November 2020



Levy Library Press *Scholarly Assessment Reports* is a peer-reviewed open access journal published by Levy Library Press. **OPEN ACCESS**