Report of the Group of Experts charged by ANVUR to advice on the process 'Valutazione della Qualità della Ricerca (VQR)'

An independent assessment on the past VQRs carried out by ANVUR

2019

Preface.

We have been invited by ANVUR, as an advisory group of experts in research production, management and assessment and in light of international best practices, to identify strengths and weaknesses of 'Valutazione Qualità della Ricerca (VQR)' assessments of Italian institutions, and suggest possible modifications or process improvements.

Following initial contacts with ANVUR in June and official nominations in July, the group met at an introductory meeting in Rome (September 14th/15th 2018), during which it decided to indicate a coordinator in the person of Professor Mauro Perretti, before deploying a road map to prepare the enclosed report.

Interaction through email and teleconferences produced a first draft of the report in October 2018 and an updated version in December 2018. Our goal was to complete the planned work with a final and undersigned report for early January 2019.

We hope that the scientific community will value our efforts, conducted in independent fashion, guided by our respect for research and research quality and integrity, and devoid of any direct or personal interest.

Overall, we find that the quality of research conducted at Italian Universities/Research Institutes, which is high on average and rather heterogeneous, can be further improved by timely and well-structured assessments, and we discuss how some key aspects of the VQR process may help towards that goal.

Faithfully

Claudio Galderisi (Hcéres and Université de Poitiers)
Mauro Perretti (Chair, Queen Mary University of London)
Nuria Sebastian Galles (Universitat Pompeu Fabra, Barcelona)
Thed van Leeuwen (Leiden University)

Rome, 12th March 2019

CONTENTS

LIST OF ACRONYMS

ANVUR. National Agency for the Evaluation of Universities and Research Institutes.

CNRS. Centre national de la recherche scientifique

CoNRS. Comité national de la recherche scientifique

DOI. Digital object identifier

GEV. Groups of experts for the Evaluation. The 16 panels of experts in the disciplines of a

scientific areas, appointed by ANVUR, that handled the evaluation of the research outputs submitted by the Institutes.

IRAS1-IRAS5. The research quality indicators by area and Institution defined by the VQR Call and calculated according to the contribution of each area to the overall value.

ISBN. International Standard Book Number

QRiH. Quality and Relevance in the Humanities

REF. Research Excellence Framework

SSD. The 370 Scientific-Disciplinary Sectors grouped into the 16 panels.

STEM. Science, Technology, Engineering and Mathematics disciplines

VQR. Evaluation of Research Quality

WOS. Web of Science

1. REMIT OF THE GROUP

1.1. The group of experts has been recruited to provide an independent assessment on the past VQRs carried out by ANVUR in order to identify strengths and potential weaknesses, in the light of other countries' experiences and best practices and to make recommendations which may be adopted for subsequent processes of evaluation.

1.2. An analysis of pros and cons on some aspects of the VQR process are presented.

1.3. The recommendations presented here are in line with our analysis and could allow ANVUR to improve the evaluation process.

1.4. In preparing and formulating our report we have read a series of documents regarding the two assessment exercises as detailed in Appendix 1. We participated in two face-to-face meetings in Rome, where we attended a presentation by the president of VQR2011-2014 and ANVUR Board of Directors, and in some other virtual meetings.

2. DEFINITIONS

The Group began its work reasoning over few fundamental considerations and definitions as presented below. Clarification on these definitions and agreement on their use and value constitute the basis onto which build the recommendations presented in Section 4.

2.1 Definition of Research

2.1.1. In this document, we define as research the intellectual, practical activities and scholarly production that researchers put in place to augment, improve and renew knowledge and, over time, ensure direct and indirect benefit and value to society. Important features of research are originality, reproducibility, common fruition and ethics.

2.2. Objectives and identified issues

2.2.1. The scientific evaluation of the products of research is a necessity that is both epistemic and ethical. As such scientific evaluation i) enables the scientific community to verify quality, relevance and impact of its hypotheses and achievements, ii) promotes scientific dialectic discussion, favouring the advancement of knowledge, and iii) guarantees the respect of the procedures of publication, valuation and disclosure of new knowledge.

2.2.2. By and large, research is funded from the public, as such it is important to have in place a process for determining appropriate and equitable distribution of public financial resources in a way that they can be spent to ensure maximal impact for Italian people. A degree of confidentiality or secrecy may be required, and would be justified, for research funded through private sources (e.g. commercial partners) or through governmental institutions (e.g. military research).

2.2.3. While not being detrimental to discovery science and the epistemic gratuitousness of fundamental research, it is important to establish the concept that publicly-funded research ought to have an impact, sooner or later, direct or indirect, and generate tangible as well as intangible values, contributing to the prosperity for the wider society, in various aspects including 'knowledge economy'.

2.2.4. The research under evaluation has been funded with public funds. In line with European Institutions, the report is an integral part of a system targeting transparency, accountability and integrity management of public funds.

2.2.5. Processes of quality assessment are not unique to Italy and take place in several countries. A very special feature of VQR is that several key aspects of the assessment methods (definition of panels, weights attached to merit classes), the definition of indicators, and their role in funding formulas are defined by laws and decrees, leaving little room for technical discretion.

The VQR evaluates the overall scientific quality of the Higher Education Institutions and Research Institutes, with the goal of linking, at least in part, public funding to quality of research deliverables. It is important to remark that the VQR assessment system focuses on research quality at the level of institutions and it does not intend to single out, either with a positive or negative accent, specific researchers.

2.2.6. Processes of evaluation of research are of importance to i) establish a snapshot on the status at a given time point, ii) provide over a period of time comparators, iii) to orient students and scientists to identify the University/Research Institutes appropriate for their development, iv) help correct any anomalies or improve research production procedures.

2.2.7. The process of evaluation of the quality of research is constructive in its nature, and not punitive. The quality and quantity of research reflects talent and efforts of researchers. Ideally, evaluation should focus not only on what is produced but also on how and why it is produced in specific institutional contexts, which may or may not efficiently use available resources and give appropriate incentives to individual and group effort. Focusing on the institutional level also reduces the relevance of causality, by the law of averages, and magnifies the role of peer pressure in eliciting efforts. A proper assessment should rely on additional relevant information on the environment in which research is performed. Alongside this perspective, we acknowledge the VQR takes into account the results of the previous assessments to place the averages of each area in their context of realization, which would make it possible to reveal the substantive trends and the progress or regressions obtained.

2.2.8. The long-term goal of the VQR (and any other research assessment process) is to elevate - over time - the quality of research produced, also in relation to the investment made by the government for publicly funded research.

2.2.9. We reason below on the pros and cons of research assessments as perceived in other countries. Brief points are listed below.

In the United Kingdom, a series of research assessment have taken place over the last twenty years, first with the Research Exercise Assessment then with the Research Excellence Framework (REF2014 to be followed by REF2021).

In favour have been i) impulse to focus and sharpen research strands within a given higher education institution; ii) as a consequence, capacity building in specific areas has occurred within research institutes and units; iii) optimisation of resources and structure reorganization to enable easy identification of research excellence from the funders and outside world in general; iv) the generation of academic posts for junior scientists (early career researchers); v) new attention to specific deliverables of research like (from 2014) its impact on society (economic, medical, societal and more); vi) the promotion of new directions (e.g. from REF2021 inter-disciplinary research is emerging as a key element). On the negative side, one could count i) the promotion of a 'transfer market' of academics with infighting between HEIs; ii) the loss of the 'academic freedom concept' or the value of science-for-the-sake-of-science; iii) the improper use of the assessment as a way to accelerate dismissals; iv) the advantaging of STEM domains over Humanities and

Social Sciences domains, even in previous to 2014 cycles of the exercise in which societal relevance was only implicitly considered as part of assessing research; v) the time-consuming efforts of UK's most prolific academics in assessing research; vi) the scale of the exercise (all universities in the UK, across all academic domains, at the same time) is considered a huge burden for the UK academic system.

Undoubtedly the processes have been successful, underpinned by an iterative refinement over the years so that at present many pros overshadow the list of cons. Mitigation rules have been inserted (for instance the uncoupling between researcher and output required by REF2021 to avoid the use of the assessment as a performance management tool). As such, the exercises will continue and will help to shape the research agenda of successful high education institutions in the United Kingdom.

In France, there is no systematic *ex-post* evaluation of research. The National Council of Universities evaluates research professors at the stage of application for promotion, research grants or sabbatical leave. Researchers employed by organizations, e.g. CNRS, are evaluated every four years by national committees (e.g. CoNRS for the CNRS). These assessments are disconnected from the evaluation of research units, which are carried out by the *Higher Council of Evaluation of Higher Education and Research (Hcéres)*. The remit of Hcéres is broad and structured to evaluate i) institutions (universities and Comue), ii) organizations courses, doctoral schools and other formative activities as well as iii) ~2,500 research units (with ~105,000 researchers). In all cases, evaluations are done in different phases. Each university or Comue submits a file and this is followed by a site visit; this process of evaluation occurs every five years by Hcéres-selected Expert Committees. These evaluations made by the expert committees do not yield a rating yet provide a platform for universities and research organizations to develop strategies, supervise research units and structure their research areas. Moreover, these evaluations take into account and favour i) coherence of the collective research programs, ii) dynamism and global influence of the research units, iii) their interaction with social, cultural and economic environment as well as iv) their national and international attractiveness. However, the absence of ratings in the evaluation reports is now questioned as currently the process does not link the evaluation report to resource allocation. This peer-review process, led by Expert Committees, is complemented by bibliometric reports produced by the Observatory of Science and Technology of Hcéres; these reports are based on the WOS for each university or Comue, organized by major disciplines. Again, these reports are not correlated with the allocation of funding to universities or Comue.

This system of evaluation has had positive impact on multi-institution programmes and it may explain the higher success rate for international projects submitted by French institutes and Institutions. On the other hand, this process of evaluation does not assess scientific contribution from each individual, nor accurately assess the scientific quality and overall impact of each institute or unit.

In the Netherlands research assessment is organized on a national scale since the early 1990's. A cyclic process, based on terms of six years, in which international peer review committees assess the quality of Dutch research, while at the three-year point, internal self-evaluation is conducted.

The system is by and large peer review based, with research metrics being an option, not an obligation (metrics are applied where considered adequate and fitting to the communication culture of the field under assessment). Fields under assessment are selected in such a way that comparison of cognate fields is avoided (for example, in a given year the assessment of physics coincides with that of law,

psychology and philosophy,). A revision of the assessment methodology in 2003 gave more autonomy to the universities, but in this manner the overarching national perspective is in part lost. A further revision in 2014 put assessment of academic quality at equal terms with societal relevance, so that Dutch academics have to argue what role they play in creating new knowledge and the way that this is transferred from academia to society at large. An important aspect of Dutch assessment procedure is the absence of a direct link between research performance assessment outcomes and funding.

2.3. Scope of the assessment and Definition of research areas

2.3.1. The task for this group is to evaluate past VQRs carried out by ANVUR. As said, VQR exercises refer to the assessment of research institutions (universities and research centres), *not individuals.* VQRs are assessments of a very large scale and the procedures have been developed to allow an evaluation process that is feasible (in time and financial means) and at the same time careful with discipline specificities.

2.3.2. Sensible definition of fields, to ensure reviewer competence, especially if products are evaluated by bibliographic metrics that vary widely between fields is of fundamental importance.

To avoid the vicious circle of arbitrary grouping/weighting/ranking by the evaluators and gaming by the evaluated, clear criteria and purposes are necessary.

2.4. Research Outputs

2.4.1. With research outputs, we indicate any scholarly product that makes an original contribution and it can take the form of monographs, book chapters, research articles, conference proceedings, software, patents and other domain-specific outputs (in particular in Arts, see below). Publications intended to *disseminate* research-based information for the use of professionals or the general public, are not considered as research outputs for the process lead by ANVUR. This is because the current assessment procedure is only considering the academic quality realm, and chooses to exclude elements that relate measurements of contributions to societal relevance.

Outputs should be identified by their DOI or ISBN, this will facilitate identification and analysis of 'multiple uses' of the same publication (otherwise difficult to control, even more if submitted by different host institutions).

2.4.2. Artistic publications refer to public outputs of artistic activities, in particular curating prestigious artistic exhibits.

2.4.3. Research publications fulfil the following characteristics

1. The generation of original data information, representing a contribution to knowledge.

2. Possibility of verification of research data information, allowing for reproducibility or verification of sources and procedures.
3. The publication takes place in channels specialised in scholarly research outputs, with editorial boards. Publications are peer-reviewed (although scholarly non-refereed publications may also be published on scholarly publication channels).

2.4.4. All research outputs are evaluated equitably across GEVs without giving an imbalance value to the outputs of different nature. Diversity and plurality of research is one of the conditions conducive to high quality and vitality of scientific research environments.

2.5. Peer-Review Process

2.5.1. The peer review description of process

Below we provide some comments, and metrics, that are referred to the VQR2011-2014.

2.5.1.1. The 2011-2014 VQR evaluation process included a bibliometric evaluation for eleven of the sixteen panels (1, 2, 3, 4, 5, 6, 7, 8b, 9, 11b, from here on indicated as 'bibliometric areas') and a peer evaluation for the other five panels (8a, 10, 11a, 12, 13, 14, from here on referred to as 'non-bibliometric areas') (See Appendix 2: List of sixteen GEV's panels).

For the bibliometric areas, an external peer evaluation was carried out on: (i) journal articles not indexed in bibliometric databases (Clarivate's WoS and Elsevier's Scopus); (ii) output typologies different from journal articles (books, book chapters); (iii) a sample of the outputs to establish the alignment between peer-review and bibliometric evaluations (on average 9.3% of total outputs, ranging from 8.4% for GEV6 to 9.9% in various GEV - see table B1, last column, in Appendix B of Final Report of VQR 2011-2014; hyperlink in Appendix 1); iv) research outputs from emerging areas at international level or in highly specialized or interdisciplinary areas.

A random sample of almost 10% (7,164 out of 77,159) outputs were assessed through both bibliometric and external peer-reviewed approaches in order to run a comparison exercise. This comparison showed an overall homogeneity of the two evaluation systems, with a global mean deviation of less than 15%. However, in some sub-GEVs the differences between the two systems were significant, close to 50%. Of particular concern is when the discrepancies resulted in 'Fair' *vs.* 'Acceptable' assessments, as they delineate the boundary between a research of quality (Fair) and a research that does not make sufficient contribution (Acceptable). It is worth noticing, that the peer evaluation tended to be more severe than the bibliometric one (Ibid., P.11, see table B5 and B6).

For all GEV Panels, the responsibility of the final scoring relied upon the Panel, through a process of informed peer review, which made use of all available information.

2.5.1.2. Peer review was also applied to journal articles when:

i) the date of publication did not adequately enable citation number to be of any relevance;
ii) bibliometric indicators - citations and journal metrics - were in conflict (sharply different from one another);
iii) the topics were considered on the margins of the relevant GEV sub-panel. Or from emerging areas at international level or interdisciplinary areas.
iv) institutions requested to do so and the GEV accepted with the request.

2.5.2. Number of peer reviewers and evaluations

2.5.1.1. In VQR 2011-2014, 12,731 peer reviewers participated in the assessment of 90,700 evaluations (Tabella 3.2 - Final Report of VQR 2011-2014; hyperlink in Appendix 1) out of a total of 118,036 research outputs which were identified.
Of note, peer-review evaluation of a relatively high number of outputs (27,938, 23.6%) could not be performed by the originally assigned reviewers (for a variety of reasons, e.g. invited reviewers did not respond; invited reviewers rejected the invitation), and were therefore reassigned to other reviewers (or were performed internally by the panel members). We note how this is not an ideal scenario, yet we appreciate the large volume of work underpinning the remaining >90,000 peer-review evaluations.

2.5.1.2. The percentage of rejected reviews is slightly higher among foreign assessors (16%: 3,358/13,719 *versus* 15%: 14,222/98,103 - see Table 3.2 of Final Report of VQR 2011-2014; hyperlink in Appendix 1).
Of the 56,448 evaluations requested for the five sectors assessed through peer review only 5,709 evaluations were conducted by peer reviewers whose institutional affiliation was abroad (11%). The percentage of outputs assessed by evaluators based in foreign institutions was higher in bibliometric areas (24%) than in non-bibliometric ones (11%). Altogether, these percentages are quite low, considering the size of the Italian research environment. Although it might be argued that Italian is the language of publication at the international level in some very specific domains, it is also true that in such domains experts from abroad are likely be able to understand Italian (for instance, specialists in Italian literature). Additionally, there is a significant number of Italian citizens working in foreign institutions who could be recruited as remote evaluators. The chances of reviewers having faced conflicts of interest are too high and it might cast some doubts about the quality of some evaluations. This is something that ANVUR will need to address in the future.

2.5.3. Average of evaluations by peer reviewer

Taking into account the figures presented in the cited Table (Table 3.2 of the Final Report VQR 2011-2014), it follows that each peer reviewer has carried out an average of seven evaluations (90,700 evaluations for 12,731 reviewers), which is compatible with the quality of an evaluation and the time it requires. Nonetheless this would depend on the type of output under assessment (e.g. book and monograph *versus* another type of scholarly publication).

2.5.4. Evaluation process

2.5.4.1. The process of peer-review required two members of the GEV panel to select one reviewer each responsible for the evaluation. When these two evaluations led to a consensus (non-disparity of merit classes), they were endorsed and validated by the GEV's sub-panel before being approved by the GEV's panel. For the VQR2011-2014, each of the 436 members of the GEV's panels had to validate on average 270 evaluations (118,036 submitted research output for 436 members). The 291 members affiliated to non-bibliometric GEV Panels (see Table 2.17 of the Final Report of VQR 2011-2014; hyperlink in Appendix 1) were asked to validate on average 295 assessments each (85,978/291). While we note the difficulty to recruit a sufficient number of peer-reviewers (as mentioned above, even more difficult to recruit reviewers affiliated to foreign institutions) we also note the sheer volume of work imposed upon the GEV Panel with a large number of evaluations which ought to be judged by each single panel member.

2.5.5. Form, disciplinary adequacy and clarity

2.5.5.1. The peer review form provides analytical judgment of a few lines and three more specific criteria: "originality", "methodological rigor" and "attested or potential impact".

2.5.5.2. Each peer reviewer indicated for each of the three criteria an annotation and proposed a class of merit. The class of merit was then confirmed (or rejected or modified) by the two members of the GEV who nominated the two peer reviewers.

3. PAST-ASSESSMENTS

We appreciate the value of the past two assessments of the status of research conducted by ANVUR (VQR2004-2010; VQR2011-2014) and in particular we praise the great effort and wealth of energies spent for VQR2011-2014. It is on the latter one that we focus our attention herein.

3.1. We have read the Report entitled *'Valutazione della Qualità della Ricerca 2011-2014 (VQR 2011-2014) Rapporto finale ANVUR 21 Febbraio 2017'* and have understood the spirit which has motored ANVUR, the Coordinator of VQR, the Chairs of the Group of Expert Valuators (GEV) panels, the chairs and members of the sub-GEVs and the broad community, making possible to conduct, and conclude, the exercise.

3.2. We recognise that the main goal of VQR2011-2014 was to assess the state-of-the-art of research in Italian Universities/Research Institutes with the medium to long term aim of favouring the production of high quality research.

3.3. We have identified many positive aspects of VQR2011-2014, which should be maintained and reinforced in future assessments. In particular, we are satisfied that VQR2011-2014 was programmed and conducted to ensure an equitable assessment as well as with a lack of non-objective influence on the final score of the outputs, and the Institutions where they were produced. Any assessment mechanism is unavoidably imprecise and incomplete in part, however, and we identify in what follows some areas of constructive improvement.

3.4. The available raw data appears to be complete and reliable, at least for research products associated to ORCID codes.

Because institutions and researchers maintain the database and select submission, their cooperation is very much needed, and may not be forthcoming if the evaluation design excites resentment. A self-confirming negative feedback mechanism can imply that lack of trust in the quality of assessment reduces that quality. In our view, the transparency of the data development process must contribute to the acceptance of results and subsequent recommendations.

3.5. There are Research Areas (GEV Panels) where clearly peer-review rather than bibliometric analyses ought to be conducted (*we provide our views on the peer-review process on 2.5).* This is a valid approach that takes into account the differences among disciplines.

3.6. The procedure used both Web of Science (WoS) by Clarivate Analytics, in conjunction with the Journal Citation Reports (JCR), parallel to the Elsevier Scopus database, in conjunction with SciVal. Citations are derived from either WoS or Scopus, Journal Impact Factors are extracted from JCR.

3.7. The bibliometric algorithm that combines these indicators is very mechanical. The choice of databases is left to the submitter, and the relative weight of journal impact factors and article cites is estimated within database journal classifications (which do not coincide with VQR assessment clusters, and may or may not be unambiguously meaningful) and item publication date.

These features enhance impartiality, but make the results somewhat opaque and hard to interpret: these mechanical criteria are quite complex, with different weights and several parameters so that the final formula may be perceived as arbitrary.

3.8. Given that evaluation is not at the individual level, but at level of the institution, the request of 2 outputs per researcher (and 3 for those employed by Research Institutes) may represent the bare minimum, considering the need to balance scientific production, high quality outputs and the need to objectively assess them within a given time limit. We recognize the need to define clearly whether the assessment is meant to detect thresholds of excellence or of minimal productivity (or both), which are essential in a publicly-funded system, and to define and possibly redefine quality and quantity thresholds in that light. In the context of massive evaluations such as the one in consideration, changes should not entail an increase of the overall number of outputs to be assessed for each research area.

3.9. The IRAS (Indicatore di Ricerca di Area nella Struttura – Index of Research Quality for Research Area in the Institution) focus on outputs (IRAS1) as well as on recruitment (IRAS2), funding (IRAS3), doctorate students (IRAS4), and institutional trajectory (IRAS5). We note that while research products are assessed on an international quality scale, all indicators are normalized, and scores essentially rank institutions on a within-field percentile scale. This may be problematic in some respects.

- Normalization within fields may usefully remove different grade inflation across fields, but deletes the absolute or relative-to-world meaning of the scores.
- Moreover, there is an issue of whether scores (computed on a global scale within WoS/Scopus subject areas) should be normalized at the level of Italian sub-disciplines (SSDs) or higher: there is a trade-off between sample size and homogeneity (not necessarily very high at SSD level).
- The resulting means and standard deviations need not be easy to interpret and precisely estimated, but should be reported if the overall assessment is homogeneous and large enough for a sensible normalization.
- Rankings put institutions against each other, and make changes over time difficult to interpret (if some improve, others must deteriorate, always along the same percentile scale). Also, improvement at the top of the scale is likely to be more challenging than improvement at the bottom (a non-linear function).

3.10. With respect to the possible corollary information, we noticed the lack of any narrative describing the research strategy, local support for research, research vision and mission, by the Italian Universities/Research Institutes assessed by the VQR2011-2014. We are aware that there is a form 'SUA-RD' which, like SUA Terza Missione, is not submitted as part of the VQR process (and only once in 5 years). Also, "Dipartimenti di Eccellenza" awards are based on VQR results (weighted by personnel in each area, as is the case at the Unit level) and do require a self-portrait and plan, which is indeed a very useful feature.

3.11. We notice the limited impact for third mission activities on the funding outcome.

4. RECOMMENDATIONS FOR FUTURE ASSESSMENTS

Below we list our recommendations as we see them conducive to replicate and if possible further elevate the quality control and the overall significance and impact of future VQR exercises.

4.1. Narrative

4.1.1. Submission to VQR should be accompanied by an Institutional narrative with on overview on the University or Research Institute and/or a more focused one at the level of Unit matching the GEV group.

4.1.2. Institutional narrative should be relative bottom-up, with little prescription, to enable freedom for each Institution to highlight their major strengths and vision within their research activities.

4.1.3. In any case, it is expected that the narrative will describe the institutional research strategy, how research support is provided, number of doctorate students, number and structure of training programmes, and third mission activities.

4.1.4. Third mission activities are important as a way to describe the impact of research, and research funding, for the further benefit of society. Disciplines in non-bibliometric areas, could consider indicators to better measure the quality of the third mission (such as done for REF in the United Kingdom, and for QRiH in the Netherlands)

4.1.5. Third mission activities would vary among specific GEVs but should include, when appropriate, social engagement that is the use of scholarly knowledge to solve societal challenges (such as commercialization, spin-outs, IP generation and management, exhibits, museums, and similar) and in general any public engagement activities. It is important the positive impact of research on third mission activities is measurable appropriately according to the disciplines.

4.1.6. A description on the strategy adopted by the Institution to maximise third mission activities, and opportunities, would be welcome and be part of the narrative.

4.1.7. The narrative document should be evaluated by the members of the GEV and could become a specific IRAS with its own weight.

4.1.8. Appendix 3 presents a proposed template for the narrative document.

4.1.9. In relation to 4.1.2, the word length of the narrative document would vary depending on the size of the Institution. Some formal structure (heading and subheadings?) of the document should be provided without being too prescriptive. This information should be given in the call and may be tailored for different GEVs.

4.2. Panels

4.2.1. VQR2011-2014 was conducted with 16 GEVs or Panels (436 members in total).

4.2.2. We recommend to increase the number of Panels to reflect better the granularity of research areas and the need to rely upon adequate expertise. Appendix 2 lists the suggested Panels (Appendix 2). One option could be to adapt to the Italian research landscape a panel organization modelled on that used at the level of the European Research Council.

4.2.3. A proper gender balance on GEV membership should remain in place, plausibly increasing the 30%:70% female to male representation for VQR2011-2014 (always taking into consideration that differences across disciplines in gender may impact the overall split). An average 10-20% increase of the female presence on the Panels would be a welcome result.

4.2.4. Actions may be taken to increase the number of GEV members working in foreign Institutions.

4.3. Bibliometric or/and peer-review analysis

4.3.1. Bibliometric analyses should be used when appropriate and fitting to the communication culture of the field under assessment. The committee is aware of the limitations of such approach, but as commented above, an individualized assessment of each output would require an extraordinary amount of resources given the magnitude of the assessment. The system deployed here takes into consideration several specificities when bibliometric analyses would not be advisable to use.

4.3.2. When bibliometric evaluation is used, it should be applied across the entire submission for that specific GEV. We suggest to specify in the call, the mechanism (bibliometric, peer-review or both) to be applied for each specific GEV. In other words, within a GEV one should work with one method.

4.3.3. One *single database* should be used for each GEV, with the suggestion that the selected database (e.g. WoS or SCOPUS) is indicated *in* the call, that is, before outputs are submitted by the Universities/Research Institutes.

4.3.4. For transparency, we suggest that single researchers should not be allowed to select a preference for bibliometric versus peer-reviewed assessment. GEVs should be able to evaluate every output in an equivalent way. While there could be distortions in very specific areas/subjects, we foresee these to be minimal since the level of evaluation is the Institution and not the individual.

4.3.5. Sampling bibliometric together with peer-review assessment should remain in place. Commendable to validate the two systems on a sample chosen randomly for the benefit of ANVUR and to ensure quality-control of the whole process.

4.3.6. We note the high impact of divisions between unclassified, D, C, B and A classes of merit for the outputs, and we are aware that their definition is out of the control of ANVUR. Nonetheless, the IRAS indicators depend heavily on the cut-offs and weights. These are written in the law and shaped in turn by the intention to use the indicators for mechanical distribution of funds, while trying not to make funding too unequal. This should be recognized but made less opaque, focusing on whether the criteria are meant to detect the extent or degree of excellency. Excellence can be used to identify and classify departments of excellence.

4.3.7. For GEVs where bibliometric analysis is applied, it is appropriate to request peer-review for borderline outputs. Borderline outputs are those falling in between or close to boundaries among unclassified, D, C, B and A outputs.

4.3.8. It is acceptable that borderline outputs are identified at the level of GEV Chair and Membership.

4.3.9 It is possible to reconsider how the definition of the boundaries between the different categories of excellence is defined and whether top 10% identifies excellence in all disciplines. It could be considered to model the boundaries for each GEV (based on specific metrics that are discipline specific). This approach would help the bibliometric review process and expedite the outcome.

4.3.10. For sub-disciplines, or sub-GEVs where citation numbers are appropriate, IRAS1 could be simplified and probably optimised by using solely citation numbers, where ranking into the top percentiles will be defined objectively for each sub-discipline or sub-GEV. The use of straight citation counts thus avoids the use of Journal Impact Factors, which are a marker of the average impact of the journal, and does not add to the impact of specific publications.

On the other hand, straight citation counts suffer from incomparability across fields, as these are not normalized by field (we note this flaw applies also on a procedure that brings together citation counts together with Journal Impact Factors as in the current IRAS1). Detailed information of evaluation methods in each GEV and sub-GEV should be provided in the call.

4.4. Transparency of the process and choice of members of GEV's panels and of peer reviewers.

4.4.1. Information on the selection process of GEV panel members and peer reviewers should be strengthened and the number of peers from foreign institutions in the different panels increased. In some panels, the percentage of 'non-Italians' is very low (less than 10%). ANVUR could adopt a mixed system - elected and appointed - for the composition of GEV panels (see 4.4.3.). Such a system is in force in other countries, notably in France, for the composition of the National Council of Universities (CNU) or for the National Committee for Scientific Research (CoNRS).

4.4.2. Some (30%-50% maximum) of the members of the GEV panel could be elected on a national basis based on their scientific reputation, though no more than 2 members from the same institution should sit on the same panel. Panel members should declare any conflict of interest as soon as they are allocated the dossiers to evaluate. They should refrain from participating in any discussion or make any comment on the submissions.

Panel members should be debriefed about ethics and other biases (including gender biases) before starting their contribution to the process.

4.4.3. As before, membership to each GEV is determined through an open call to ensure transparency, equality and diversity. However, this approach could be complemented by ad-hoc invitations by a central Panel (for instance composed by the GEV chairs and sub-chairs). In this manner, this mixed approach with the directed intervention by a central Panel will ensure equality and diversity, with proper distribution with respect to gender, sub-discipline and geography. The relative proportion of the two ways for recruiting could be decided by ANVUR or others, e.g. 70/30 for open call *vs.* direct nomination or 50/50 or else, keeping in mind that excellence and quality are the main determinants.

4.5. Terms of the evaluation process

Given the number of members of the GEV's panels (27 on average per panel) and the number of evaluations per area (on average 7,300), it seems reasonable to limit the validation process by the two GEV's members in charge of manifest cases of disparity of judgment (difference of two classes of merit). In this case, a third external expertise to the GEV panel appears necessary.

4.6. Peer-review evaluation form

4.6.1. The three criteria used for the peer-review process applied within VQR2011-2014 probably do not fully reflect the qualities required for a publication of excellence. The third criterion ('attested or potential impact'), where the research has been exerted, or is likely to be exerted in the future, has a theoretical and/or applied influence.

Such a criterion brings together two rather different criteria that are difficult to evaluate objectively. The adjective 'attested' seems to refer to quantitative or bibliometric values, whereas the 'potential' character of an impact seems difficult to evaluate, and introduces a degree of subjectivity that can hinder the understanding and acceptance by the scientific community of the evaluation process. In short, we see a degree of contradiction of terms herein.

4.6.2. The peer reviewers should assess each criterion with a rating that ranges from 0 to 8 points (Excellent=8; Good=6; Acceptable=4; Limited or Insufficient=2; inadequate=0). In any case, ANVUR should provide a scale indicating the correspondence between numerical grades and classes of merit. It would be advisable to provide reviewers with examples and clear guidelines. For instance, at European Research Council level, Panel Members are informed about the data produced in previous evaluations, providing in this manner a helpful background.

4.6.3. ANVUR should present peer-reviewers with a more articulated form and for instance five criteria instead of the current three. An example of this new expansion of the criteria could be: a) methodological rigor, b) knowledge of the state of the art; c) originality, d) use of acquired knowledge; e) quality and clarity of the argument and the purpose.

An output that scored five times the Excellent score would then have 40 points, whereas a product that scored two Excellent and three Good would be awarded an overall score of 34.

This would avoid the effects of thresholds of the current system, decrease the share of subjectivity and hopefully would mitigate litigation.

4.7. Outputs

In section 2.4 we indicate our definition of research output. Similarly, we have defined the features of research in section 2.4.3. The process of VQR intends to assess the quality of the research and its outputs.

In line with VQR2011-2014, where there was an adherence of 98%, each researcher submitted by an Institution ought to be identified by an ORCID number.

4.7.1. Researchers should submit 2 or 3 outputs when employed by Universities or Research Institutes, respectively. These numbers would be a reasonable requirement to an active academic over several years of assessment period. We note that affiliation at time of assessment, rather than of acceptance or publication, makes products acceptable. This potentially penalizes previous investments at the institutional level because researchers' mobility is not symmetrical. Indeed, investigators tend to move to better research institutions (in particular young researchers) inside and outside Italy.

ANVUR should consider the possibility of institutions submitting a number of outputs (5-10% of the total outputs) by researchers no longer affiliated with them. Although the mobility of researchers in Italy is low, the impact may be significant in some departments who cannot retain (or attract) excellent investigators. This proposal would complement the IRAS2 indicator of performance through staff promotion or recruitment.

4.7.2. Another option to consider is to broaden the tools available for the assessment so that each researcher is allowed to submit a number of publications ranging from 1 to 5, which would make it possible to better measure the overall quality of the Research Unit or Department. However, this would increase the overall volume of the assessment for the members of the GEVs and the need to identify a larger number of peer-reviewers and/or evaluators. In any case, and to avoid confusion, the number of outputs should be fixed at the time of call for each GEV.

4.7.3. It is fundamental that the outputs reflect the true contribution of the researcher is accepted and implemented across the entire process. The following measures should be in place to properly assess researchers' contribution in each output and appear in the call:

a) Based on the VQR2011-2014 data (pirate plots), for each GEV and sub-GEV a pre-specified number of authors per output will be determined.

b) For outputs with less than the pre-specified number of authors, the contribution will have the same weight.

c) For outputs with more than the pre-specified number of authors (see 4.7.3a), appropriate weighting will be assigned depending on the position of the authors in the list of authors, or following what is customary in the field. For instance, for life sciences, first and last authors will receive full weight. In many other disciplines, corresponding authors could receive full weight.

4.7.4. The number of pre-specified authors and the weighting system for the each GEV and sub-GEV should appear in the call. In fact, we note that in some disciplines authorship follows alphabetical order (e.g., mathematics, economics), while other fields have clearly defined work flows in which authorship is organized by convention (e.g., high energy physics). Each GEV and sub-GEV should have the freedom to determine on issues concerning authorship when in the evaluation of an output differences with the pre-specified parameters are encountered.

4.7.5. Researchers and institutions have to consider the impact of the authorship weighting system for the selection of their outputs. Therefore, if information about authors' contribution is not available, it may be preferable to select outputs where the contribution of the researcher is clearly stated (e.g. in the Authors Contribution section, wherever available).

4.7.6. Clear information about the assessment of researchers' contribution to the outputs should be available in the call for outputs that are not publications (such as software or coordination of research expeditions, for instance). These could vary depending on the respective GEV.

4.7.7. In most disciplines review articles should not be considered as an original piece of research. Again, this could vary depending on the respective GEV and ought to be specified in the call.

4.8. Assessment

4.8.1. For bibliometric analyses, the percentages associated with each band should remain as in VQR2011-2014.

4.8.2. We propose to apply the same IRAS as for VQR2011-2014, with minor modifications.

4.8.3. IRAS1 should remain as for VQR2011-2014 with a weight between 60 and 70%. Alternatively, ANVUR should consider the possibility to modify this IRAS as per clause 4.3.10.

4.8.4. IRAS2 is an important score, as it reflects the tangible activities within each Institution to deploy their research strategy. It also provides a tangible metric on research support and planning. Its weight should be between 10 and 20%.

4.8.5. IRAS3 should remain as it is as it reflects the ability of Institutions to win competitive funds for research. Its weight should be between 3 and 5%, with the higher banding perhaps assigned over a specified threshold for funds awarded through commercial sources.

4.8.6. IRAS4 and IRAS5 should be merged into a new one (new IRAS4) resulting from the evaluation of the narrative document, with a weight between 5 and 10%.

4.8.7. Doctoral students and training programmes, together with relative metrics, will be included in the narrative document (new IRAS4).

4.8.9. Appendix 4 presents potential models for the 4 IRAS and their respective weights.

4.8.10. Furthermore, ANVUR could consider to stagger the assessment process and evaluate specific disciplines at different times, e.g. humanities separated from STEMs. This would distribute the burden of the assessment over the years and would impede non-comparable comparison. On the negative side, the burden is spread out over years and might produce a feeling that assessment takes place continuously.

5. CONCLUSIONS

5.1. A fair evaluation needs respected evaluators and clear assessment criteria. It can have multiple objectives and criteria but it should avoid any sense of arbitrariness. An atmosphere of distrust and hostility is not conducive to a useful evaluation, and the issue is particularly relevant in academic environments that find it hard to achieve reciprocal and societal respect.

5.2. The Research Quality Assessment (VQR) process implicitly entails that ANVUR continues to adhere to these requirements and criteria. It will be important to build strong adherence to the declarations of San Francisco and Leiden in the process of evaluation.

5.3. The analyses and recommendations presented here aim to improve transparency, effectiveness and acceptability of the overall evaluation.

5.4. The experts hope that the present document contributes to fostering dialogue and trust within the Italian scientific community, as a condition of a shared awareness of the scientific and ethical benefits that derive from an evaluation of serious and conscientious research.

APPENDIX 1 – documents assessed

A) Call of VQR 2004-2010

B) Ministerial Decree n.17of July 15, 2011

C) Group of Experts' criteria of evaluation for VQR 2004-2010

D) Final Report of VQR 2004-2010

E) Ministerial Decree n. 458 of June 27 2015

F) Call of VQR 2011-2014

G) Call to serve in the Panels of VQR 2011-2014

H) Report on Group of Expert selection

I) Comments to provisional Call of VQR 2011-2014

L) Group of Experts' criteria of evaluation for VQR 2011-2014

M) Final Report of VQR 2011-2014

N) Third Mission Evaluation Manual

APPENDIX 2 - List of sixteen GEV's panels

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 8a Architecture

Area 8b Civil Engineering

Area 9 Industrial and Information Engineering

Area 10 Ancient History, Philology, Literature and Art History

Area 11a History, Philosophy, Pedagogy

Area 11b Psychology

Area 12 Law

Area 13 Economics and Statistics

Area 14 Political and Social Sciences

APPENDIX 3 – draft template for the narrative document (IRAS4)

Submission to VQR should be accompanied by an Institutional narrative with on overview on the University or Research Institute and/or a more focused one at the level of Unit matching the GEV group.

Potential Sub-headings:

1. Describe your research achievements and strategy in the period under assessment, and how the Institution has supported it.
2. Describe your desired research goals in 5-year time and the strategy of the Institution to deliver them.
3. Describe the activities of the Institution towards internationalization and strategic partnerships.
4. Describe the institution approach to assess and support the broad impact of their research.
   (Third mission activities would vary among specific GEVs but could include, when appropriate, social engagement that is the use of scholarly knowledge to solve societal challenges (such as commercialization, spin-outs, IP generation and management, exhibits, museums, and similar) and in general any public engagement activities.)

   To help completion of IRAS4, the information contained Scheda SUA-RD and SUA-TM could be used.

APPENDIX 4 - IRAS and their potential specific weight.

IRAS1: as for VQR2011-2014, with a weight between 60% and 70%

IRAS2: as for VQR2011-2014, with a weight between 10% and 20%

IRAS3: as for VQR2011-2014, with a weight between 3% and 5%

IRAS4: NARRATIVE DOCUMENT with weight between 5% and 10% (see Appendix 3)