2023

VQR 2015-19

EXPERT REVIEW PANEL

JUNE 2023

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The *Valutazione Qualità della Ricerca* (VQR) is Italy's national research evaluation. The most recent evaluation was issued in July 2022 and considered research published between 2015 and 2019. In January 2023, ANVUR invited a panel of international experts to review the most recent VQR 2015-19. The expert panel was provided with information about the process of the evaluation, alongside detailed analysis of VQR 2015-19 conducted by ANVUR. This report presents the considerations of the expert panel along with their recommendations.

PRINCIPAL AREAS OF FOCUS

In making recommendations, the panel has focussed on a number of key areas in relation to the VQR, learning from the previous assessment to inform the development of future assessments. The areas of focus are:

- **Purpose(s) of the assessment.** Clarity about the purposes of national research assessment is a central consideration that has implications for all decisions about the assessment. The panel agreed that the primary purpose of the VQR is clear, in that the exercise is conducted to inform allocation of funding. However, the panel noted other potential purposes and implications that merit further consideration.

- **Structure and scale of the assessment.** The panel considered the structure and scale of the assessment and suggested that there was potential to reduce the scale of parts of the assessment without impacting its robustness. The panel also considered options for broadening the VQR to include more aspects of an excellent research environment.

- **Panel recruitment and use of external experts.** The panel considers that peer review by experts should continue to be a central part of the assessment. As a result, the selection of panel members and reviewers is a critical success factor for the exercise. The panel makes a number of recommendations in this area including reforming the selection of panel members, the use of external experts, and the training of panel members and reviewers. In the medium term, the panel also considered that the use of artificial intelligence to match outputs to reviewers should be explored in due time.

- **Use of bibliometrics.** The panel reviewed the current use of bibliometric indicators to support the assessment of outputs, and emphasised the importance of international good practice in this area which is evolving rapidly. While some

indicators should not be used (journal-level metrics and altmetrics), any use of other indicators should be accompanied by appropriate training for reviewers.

- **Assessment of Social Sciences, Arts and Humanities.** The panel considered that the assessment of research in social sciences, arts and humanities deserves special attention. In particular, the increased diversity of research outputs in these disciplines should be taken into account in the assessment, as well as the methodological debates that can be important factors in research quality.

- **Open research.** The panel considers that increasing open access to research outputs, and the increasing adoption of broader open research practices are important features in the research system. The VQR has a potential role in supporting and incentivising open research practices.

- **Third Mission.** The panel commends the increased focus on Third Mission in the recent iterations of the VQR. Measuring the benefits to society that are created by research organisations has an important role to play and the VQR has developed a robust methodology that is in line with international good practice. The panel considers that an increasing focus on Third Mission assessment should be developed in the future VQR, including consideration of whether more funding should be allocated based on this part of the assessment.

SUMMARY OF RECOMMENDATIONS

Recommendation 1: The panel recommends that the Ministry of Education and Research in collaboration with ANVUR publishes a clear statement of the purposes of the VQR. This should include those additional purposes beyond the allocation of funding.

Recommendation 2: The panel advises augmenting the existing elements to achieve a more holistic assessment of research performance in universities and other research organizations, aligning with national priorities.

Recommendation 3: The panel recommends reviewing the scale of the VQR, with a view to reducing the number of research outputs that are assessed.

Recommendation 4: The panel recommends strengthening the effectiveness of the evaluation process by implementing targeted measures to enhance reviewer selection. Panel chairs of the "Gruppi di Esperti della Valutazione" (GEV) should be empowered to actively recruit GEV members and external reviewers, enabling them to identify individuals with specific qualifications and expertise with ANVUR support. While continuing to consider the

diversity of GEV membership, the practice of selecting members by drawing lots should be removed from the process.

Recommendation 5: ANVUR should foster diversity and expertise within the reviewer pool by proactively increasing the percentage of external reviewers per output involved in the evaluation process, thereby increasing the chances of having a panel that effectively covers all the competences needed.

Recommendation 6: ANVUR should establish ad hoc training sessions for panel chairs and vice-chairs to ensure that the purpose, procedures, and desired output of the evaluation is well understood. If possible ANVUR should assign an officer to each panel who could support and advise on the conduct of the panel's work, to ensure consistency among panel members.

Recommendation 7: The scoring across the exercise needs to be normalised according to disciplinary practices. In the higher education system, some institutions focus on disciplines that traditionally show lower publication patterns or that do not commonly use metrics in assessment. Such institutions are therefore penalised when there are no normalisation processes applied to reflect the different approaches across and between disciplinary areas in the evaluation.

Recommendation 8: Over time, ANVUR should carefully explore the use of AI tools to facilitate the assignment of experts to research outputs. While recognizing the value of partial automation, it is essential that GEV panel chairs and vice-chairs retain the autonomy and opportunity to review such mechanisms.

Recommendation 9: The panel recommends that ANVUR explores the possible use of existing platforms, like Clarivate Reviewer Locator, in the first place. Alternatively, a dedicated platform based on Natural Language Processing (NLP) can be developed over time with the appropriate support of the Minister. Both options will require a dedicated effort and can be considered for the future.

Recommendation 10: Journal-level metrics should not be used as part of the assessment in the VQR in the future. ANVUR could select other indicators of their choice and ensure that all panels use them with the same criteria, possibly also in consultation with chair and vice chairs of panels from previous evaluations.

Recommendation 11: ANVUR should ensure that GEV members and external reviewers receive specific instructions on the use of bibliometrics and proper training to avoid the inappropriate use of some metrics.

Recommendation 12: ANVUR should not incorporate the use of altmetrics into the next VQR.

Recommendation 13: ANVUR should ensure that experts in research methods are included in the evaluation process either for briefing and/or within disciplinary GEVs.

Recommendation 14: ANVUR should retain the double scoring of monographs compared to journal articles, regardless of their distribution. It should also encourage submission of different research outputs beyond publications through the ANVUR guidelines but leave the assessment to the panel members' discretion.

Recommendation 15: ANVUR should encourage a fair evaluation of research outputs regardless of the language in which they are produced.

Recommendation 16: ANVUR should consider introducing upper limits for the percentage of non-Open Access product and that data generated with public funds are made available in accessible repositories. Finally, as different disciplines have a different approach to OA publications, ANVUR should provide guidelines around assessment and scoring of OA publications.

Recommendation 17: The panel recommends that consideration is given to allocate an increased proportion of funding on the basis of the assessment of Third Mission.

Recommendation 18: The panel recommends that ANVUR continue to align their approach to Third Mission assessment with EU recommendations and practices in Australia, Hong Kong and UK where appropriate.

Recommendation 19: The panel recommends that ANVUR reviews the fields of action for Third Mission case studies and considers a more flexible approach to classifying case studies.

Recommendation 20: The panel recommends that guidance is provided to ensure consistent scoring of Third Mission case study between the fields of action.

Recommendation 21: The panel recommends that in future Third Mission assessments, the number of case studies required should be directly linked to the size of institutions.

INTRODUCTION

In common with some other nations, over recent years Italy has conducted a national evaluation of the research performed in its universities and other research organisations. The *Valutazione Qualità della Ricerca* (VQR) has been conducted three times, with the most recent evaluation being published in July 2022 and considering research published between 2015 and 2019. The evaluation is conducted by the *Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca* (ANVUR).

In January 2023, ANVUR invited a panel of international experts to review the most recent VQR 2015-19 and make recommendations concerning the development of the VQR in the future. The membership of the panel is provided in the Annex. This report presents the findings and recommendations of the expert review panel.

BACKGROUND ON VQR 2015-19

The VQR 2015-2019 evaluated the results of the scientific research of Italian Institutions and related internal divisions (departments and similar units), also taking into account the scientific area. The exercise also aimed at evaluating Third Mission activities carried out by the Italian Institutions and their internal divisions, the impact of which occurred in the period 2015-2019.

The results of this exercise are used by the Ministry of Universities and Research:

- for allocating the performance-based share (the Quota premiale) of the main university funding (FFO, Ordinary Financing Fund; for 2022 the premium share is EUR 2.336 billion, almost 30% of the total fund)
- for the allocation of the share of the FFO earmarked for the co-financing of research projects submitted by Inter-University Research Consortia (for 2022 it amounts to EUR 2 million)
- for selecting the 180 excellent departments of Italian Universities that will obtain extra financial support for 5 years (the "Dipartimenti di Eccellenza" Program, varying between EUR 1,620 and 1,080 million annually for five years).

The new VQR had novel elements compared to the previous evaluation exercises (periods 2004-2010, and 2011-2014). Although it maintained the basic methodological principles of peer evaluation, also informed by metrics where appropriate, and an assessment based on originality, rigour and impact, innovative elements affected mainly on expanding the scope and scale of the evaluation.

A more innovative approach was also addressed to the recruitment and selection procedures of the Evaluation Experts Group (Gruppi di Esperti della Valutazione, GEV) which included a separate panel for the assessment of impact/interdisciplinary or Third Mission case studies submissions, adding extra 30 members to the overall number of disciplinary experts, and the review of the disciplinary panels to accommodate emerging studies particularly in economics and statistics, and business studies and statistics.

Assessment categories were also revised and recalibrated, compared to previous exercises. Yet, this remains an area of refinement as ANVUR wishes to identify a more continuous indicator than the current assessment categories. Furthermore, a significant effort was dedicated to the calculation of quality profiles for permanent staff and those recruited or promoted during the period of the evaluation, 2015-2019. For the latter, particular attention has been paid to the university awarding the PhD degree rather than those where the researcher is currently affiliated. This is of particular relevance to assess training capacity of universities across different regions, in a country that shows a wide gap of overall performance among universities.

Finally, a quality profile was also included for Third Mission activities. It has been introduced for the very first time to ensure a more thorough evaluation of exploitation and dissemination of research in harmony with emerging global standards.

INTERNATIONAL CONTEXT

There is considerable interest and debate on the processes and impacts of research assessment at an international level; the growing acknowledgement that national research assessments (alongside other contexts where research is evaluated) create incentives within the research system that can have both positive and negative consequences. Over the last decade, there have been a number of statements and initiatives aimed at improving the practices of research assessment, starting with the San Francisco Declaration on Research Assessment (DORA)[1]. Key contributions to this debate include the Leiden Manifesto[2], The

[1] https://sfdora.org/

[2] Hicks, D., Wouters, P., Waltman, L. et al. Bibliometrics: The Leiden Manifesto for research metrics. Nature 520, 429–431 (2015). https://doi.org/10.1038/520429a

Metric Tide report[3], ongoing work in the context of the Global Research Council[4], and, most recently, the Coalition for Advancing Research Assessment (CoARA)[5]. All these initiatives reflect a common concern: there is a risk that research assessment, especially using mainly bibliometric indicators, can take an overly narrow view of research excellence and that a move to more holistic, contextualised assessment should be encouraged.

REMIT OF THE EXPERT GROUP

The panel of experts for the review of the evaluation was appointed in March 2023 by the Governing Board of ANVUR. The panel had a clear mandate to prepare a report assessing strengths and weaknesses of the VQR 2015-2019 conducted by ANVUR in consideration of international best practices. Also, the report must include a preparatory work to launch the next evaluation exercise, that will cover the period 2020-2024, and propose criteria and methodologies for the next VQR exercise to provide prompt guidance to the participating institutions.

The Agency provided a "Critical Analysis of the VQR 2015-2019 Results" prepared by the two units of research quality and Third mission assessment. The report is a complete analysis of the data of the evaluation and has been the main resource for the assessment of the panel. At a meeting organised by ANVUR in Rome on March 22nd, 2023, the management team provided a set of presentations for the data and listed the key areas of analysis for which the agency wanted specific feedback.

In particular, they highlighted the following issues:

1) Informed peer-review and bibliometric indicators.
2) Assessment in the SSH fields.
3) Experts database and reviewers' management.
4) Evaluation criteria for disciplinary fields, Open Science and possible assessment of infrastructure.
5) Assignment of scores and the transcoding table.
6) Evaluation criteria for Third mission activities and fields of action.

[3] Wilsdon, James; Allen, Liz; Belfiore, Eleonora; Campbell, Philip; Curry, Stephen; Hill, Steven; et al. (2015). The metric tide: report of the independent review of the role of metrics in research assessment and management. University of Sussex. Report. https://hdl.handle.net/10779/uos.23418680.v1
[4] https://globalresearchcouncil.org/about/responsible-research-assessment-working-group/
[5] https://coara.eu/

Beyond the wealth of data provided, that is the report with the analysis of the result and the two sets of slides, the panel also requested a further analysis with a simulation considering lowering the output requirements per researcher.

All the documents have been carefully analysed and will be cited and mentioned in this report. The remainder of this report is structured around a set of key areas of observations and recommendations from the expert panel and guided by their knowledge and expertise. The panel suggests that the data accessible through CINECA, along with the evaluation data encompassed in this report, could be made readily available for the advantage of higher education institutions and research centres.

PURPOSE(S) OF THE VQR ASSESSMENT

The purpose of any assessment process should be clearly understood by the participating entities. In this particular case, both the national authorities and the participant institutions need to be fully aware of the final goal of the evaluation: what are the reasons, how will the results be used and what are the consequences of the assessment for universities and research institutes. The panel has formed the view that these would benefit from further development, and consideration of how the 'purpose' of the exercise is communicated to the participating organisations as well as to the evaluating panels.

For the national research authorities, the panel would also highlight the importance of a clearly defined purpose and process for the exercise. National research authorities that initiate research assessments benefit from clearly defining their motivation beyond the fair distribution of research funding between national universities and research centres. The understanding and an internal consensus on purpose(s) improves the evaluation design and the downstream use of the results for resource allocation.

The panel also formed the view that a clear communication of the purpose can provide incentives for participating organisations to ensure the quality and richness of the data submitted. All participants should have a clear understanding of the benefits of the outcomes of the process for institutions, alongside any statutory obligations the participants may have.

The panel acknowledged that for many assessment processes, the primary purpose can be to provide evidence to underpin the national research resource allocation. However, there are other dimensions to the 'health' and performance of a national research system that can be tackled through a national evaluation process. As lessons from other countries suggest, these dimension could include improving the overall quality of research, encouraging wider

participation in research activities, developing academic careers and institutions, increasing international visibility of research, improving success rates in applying for external funding, supporting, coordinating and improving national infrastructure, encouraging quality in research training (Ph.D), and increasing collaboration with public and private organisations to improve the benefits that the economy and wider society derives from the initial research investment.[6]

The panel also highlights that having a clearly expressed purpose can provide some assurance that the process will not create a set of unintended consequences or negative incentives which undermine the operation of the process or trust in the process.

Recommendation 1: The panel recommends that the Ministry of University and Research and ANVUR publish a clear statement of the purposes of the VQR. This should include those additional purposes beyond the allocation of funding.

STRUCTURE AND SCALE OF THE ASSESSMENT

As mentioned above, the innovative elements of the new VQR methodological approach impacted mainly on the structure and scale of the exercise. In fact, the division into five instead of four years has increased the submission of peer-researcher output from two to three, and the submission of research institutes and universities has been the same.

Another main characteristic of this exercise was the flexibility in the number of outputs of individual researchers, from zero to four as long as the overall submission of a department remained three times the number of its researchers. In fact, the aim of this change clearly suggests that ANVUR targeted the overall performance of a department, rather than assessing individuals. Individual scoring should also be avoided when the assessment is addressed to understand institutions' performance. The new call also included the requirement to introduce additional products, if the same product was presented by a number of institutions higher than the threshold set by the call.

---

[6] See The changing role of funders in responsible research assessment: progress, obstacles and the way ahead (RoRI Working Paper No.3); also see Hicks, D. (2010). Overview of models of performance-based research funding systems. In *Performance-based Funding for Public Research in Tertiary Education Institutions* (pp. 23-52). OECD. http://doi.org/10.1787/9789264094611-4-en; See Ochsner, M & Peruginelli, G (2021) *National Research Evaluation Systems and the Social Sciences* in Engels, T. C. E. & Kulczycki, E. (Eds.) *Handbook on Research Assessment in the Social Sciences*. Edward Elgar, ISBN 9781800372542. https://serval.unil.ch/resource/serval:BIB_F8DED5348DC3.P001/REF.pdf

The VQR 2015-19 assessment considered evidence in two areas: research outputs and third mission case studies. This provides a particular focus and the panel notes that the data on research outputs is also used to determine the quality of research training in the institutions assessed. However, this focussed use of evidence omits some other features related to excellent research performance. Examples include:

- the hosting of high-quality research infrastructures[7], including infrastructures that contribute to the pursuit of research outside the institution itself at the national or international level;
- the number of and success rates in obtaining competitive national and international funds;
- the number and quality of PhD students graduated and post-doctoral research associates supported, and support for their future career progression;
- support for open science practices, including open access publications, open data and open software code;
- contribution to peer review of grants and publications at a national and international level;
- good practices concerning research integrity and reproducibility.

The absence of these features from the assessment has implications on the incentives for improvement for institutions and departments. For example, external funding serves as a catalyst for innovation and showcases researchers' ability to secure crucial financial support. Meanwhile, Ph.D supervision fosters academic mentorship and cultivates the growth of future research leaders. Recognizing the significance of these contributions aligns with international standards and promotes a comprehensive evaluation framework that enhances the potential for impactful research outcomes.

As far as research infrastructures are concerned, there are two classes that could be considered. The first one covers infrastructures that are scientific achievements themselves. They can be included as research outputs in the evaluation if they are a subject of a publication in a monograph, journal, book, conference proceeding or a patent. It is not clear, however, if infrastructures that were not published in this way can qualify as "Other scientific outputs" (defined in Art. 5.2.e of the Presidential Decree no. 1 of January 3rd, 2020).

[7] For a definition: Research Infrastructures (RIs) are long-term enterprises, often dynamically operating for several decades. They represent strategic investments which are indispensable for enabling and developing research in many scientific domains and play a major role in innovation and science. See ESFRI definition https://www.esfri.eu/research-infrastructure-ri; also see Hallonsten, O. (2012) *Continuity and Change in the Politics of European Scientific Collaboration* in Journal of contemporary European Research, 8 (3); (2014) *The Politics of European collaboration in big science*, The Global Politics of Science and Technology-Vol. 2: Perspectives, Cases and Methods, Springer Berlin Heidelberg, pp.31-46; Jacob, M, Hallosten, O. (2014) *The persistence of big science and mega science in research and innovation policy*, Science and Public Policy 39, (4) pp. 411-415. See also the ESFRI Landscape Analysis (2024) in preparation.

Research infrastructures that serve as enablers of experimental research play a crucial role in shaping outcomes. It is important to acknowledge their significance and explore effective ways to include them in the evaluation process in the long run. These infrastructures contribute significantly to the overall research environment and can greatly impact the quality and novelty of scientific advancements. Therefore, it is advisable to consider suitable approaches for their inclusion in a comprehensive manner.

Determining the prioritisation of these areas, if any, is a matter of national policy. It may not be feasible or desirable to incorporate all of the suggested aspects mentioned above. Furthermore, it is crucial to carefully examine potential unintended consequences. For instance, if the assessment includes PhD student numbers, there is a risk of departments over-recruiting at the expense of the quality of training provided. It is equally important to present any additional quantitative metrics for research performance within their appropriate context and normalise them to account for disciplinary variations. In essence, the panel emphasised the need to consider whether a more comprehensive and holistic evaluation of the research environment in departments should be conducted alongside the ongoing assessment of research outputs and third mission activities.

Recommendation 2: The panel advises augmenting the existing elements to achieve a more holistic assessment of research performance in universities and other research organisations, aligning with national priorities.

The VQR is a large-scale exercise, with the assessment of research outputs requiring a considerable amount of effort. In VQR 2015-19 each researcher was required on average to submit three research outputs for assessment, resulting in 182,648 outputs being assessed. In addition to the 615 GEV members being involved in peer review, over 11,000 external reviewers were also involved. The VQR is to be commended for its thoroughness and commitment to robust outcomes. However, the panel notes that the number of outputs assessed is high. The VQR 2015-2019 covers a five-year period, so the output requirement is 0.6 per researcher per annum. This contrasts, for example, with REF 2021 in the UK, where 2.5 outputs per researcher were required to cover a seven-year period, a requirement of 0.36 per researcher per annum.

Following a comprehensive analysis conducted by ANVUR at the panel's request, it has been determined that the current VQR can effectively adjust the size and scale of the exercise without compromising the robustness of the assessment. The analysis reveals that even with a reduction in the number of research products considered, the assessment results will remain reliable. These insights indicate that the VQR can adopt a more streamlined approach while maintaining the integrity and accuracy of the assessment process.

Recommendation 3: The panel recommends reviewing the scale of the VQR, with a view to potentially reducing the number of research outputs that are assessed.

PANEL RECRUITMENT AND USE OF EXTERNAL EXPERTS

In the VQR 2015-19, Expert Groups (GEV) recruitment was another area of innovation. As mentioned above the number of experts was enlarged to include a separate panel dedicated to interdisciplinarity, impact, and Third Mission and additional panel members when necessary and if requested by the GEV panels. The evaluation of research outputs involved a meticulous process of assessment by reviewers from the GEV and additional experts. These additional reviewers, consisting of Italian and international scholars, were carefully selected based on principles of cooperation, objectivity, impartiality, and correctness. The GEV members, on the other hand, choose reviewers from a list provided by ANVUR, ensuring alignment of expertise with the research topics at hand. The CINECA.IT platform was utilised to verify conflicts of interest, examining institutional affiliations and co-authorships during the reviewer selection process.

In total, 645 experts were actively engaged in the VQR exercise including 30 for a dedicated panel to Third Mission, interdisciplinarity and impact; over 11,000 additional reviewers requested by the GEVs also participated in the evaluation. The panel commends the ANVUR for their efficiency in selecting reviewers and for the appropriate performance of their digital platform. The Agency's diligent approach of choosing qualified reviewers and the reliable functionality of the digital platform significantly contributed to the overall success of the evaluation process. Their dedicated efforts ensured a smooth workflow and optimal resource utilisation. The overall outcome of the review reflected a satisfactory level of quality, underscoring the organisation's ability to adeptly handle the task and deliver commendable results.

Peer review assessment remains the gold standard for national research evaluations. To uphold international standards and ensure the selection of qualified experts, agencies conducting evaluations must meticulously choose reviewers, maintain and retain a robust pool of expertise, and encourage participation from both national and international evaluators.

However, the evaluation process within the existing structure faces various challenges that demand attention. These challenges include:

- the complexity of the evaluation process itself
- the selection of highly qualified reviewers
- addressing potential biases and conflicts of interest
- maintaining consistency in scoring among reviewers

- the imperative need for diversity in reviewer selection
- and the integration of AI tools.

Overcoming these challenges is crucial to enhance the integrity, reliability, and fairness of research evaluations within the national context and beyond. The selection of GEV members by drawing lots, randomly, from a list of candidates who met scientific requirements, was newly introduced in this exercise to guarantee a fair and balanced representation of gender, academic position and disciplinary composition. A further check was secured to prevent conflict of interest of the members of the evaluation panels. The attention of the Agency is commendable yet, drawing lots entails the risk of limiting the panel's academic expertise and, therefore, the fairness of the evaluation process, especially in panels where specific competences need to be prioritised.

After careful deliberation, the panel has put forward recommendations aimed at improving the evaluation process conducted by ANVUR. These recommendations address key aspects such as the required number of outputs and reviewer allocation. The reduction in the required number of outputs (see recommendation 3, above) and an increased involvement and diversity of the external reviewers, should enhance the quality and inclusivity of the evaluation process. Implementing these recommendations will further strengthen ANVUR's evaluation framework, fostering improved outcomes and instilling greater confidence within the academic community.

It is therefore suggested that more autonomy could be secured to the GEV panel chairs and vice-chairs in the management and composition of the panel so that specific qualification and expertise could be secured. Panel chair and vice-chair must be supported by automated tools and large databases of relevant experts and by the ANVUR personnel.

It will remain mandate to the Agency to secure a fair and balanced representation of panel members once the first selection has been conducted by chairs/vice-chair only on scientific merits and criteria.

Finally, briefing sessions for all reviewers is a fundamental step to ensure all GEV members and external reviewers are clear about the purpose(s) of the evaluation. As this is a fast-changing environment, it is worth considering including in the training both mainstream as well as distinct methodological overviews of what quality means in different disciplines, which will encourage consistency of assessment among different reviewers. It may be helpful to allow experts with more methodological training in the disciplines covered by the panels. Research on scientific methods and disciplinary cultures are emerging as strong and important elements of assessment in research design and implementation; and more scholars per different disciplines are focusing on practices of research and "research on research" studies. It is crucial to raise awareness and grow expertise and capacity for all evaluators in disciplinary and interdisciplinary contexts. Methodological experts could

either provide briefing at the start of the evaluation process and also be actively involved in the evaluation from within the disciplinary panels.

It is understood that such training requires considerable effort and resources on the side of ANVUR. To mitigate the effort, it is advisable to retain a fairly large group of experts from a previous VQR to the next, taking care of sufficient diversity of such groups. Training and guidance of experts already involved in earlier VQRs would be simpler and shorter than training and guidance of experts that are involved in VQR expertise for the first time.

Recommendation 4: The panel recommends strengthening the effectiveness of the evaluation process by implementing targeted measures to enhance reviewer selection. GEV panel chairs should be empowered to actively recruit GEV members and external reviewers, enabling them to identify individuals with specific qualifications and expertise with ANVUR support. While continuing to consider the diversity of GEV membership, the practice of selecting members by drawing lots should be removed from the process.

Recommendation 5: ANVUR should foster diversity and expertise within the reviewer pool by proactively increasing the percentage of external reviewers per output involved in the evaluation process, thereby increasing the chances of having a panel that effectively covers all the competences needed.

Recommendation 6: ANVUR should establish ad hoc training sessions for panel chairs and vice-chair to ensure that the purpose, procedures, and desired output of the evaluation is well understood. If possible ANVUR should assign an officer to each panel who could oversee the correct development of the panel and ensure homogeneity among them.

These steps serve multiple purposes: they minimise the potential for bias, mitigates conflicts of interest, and avoid undue concentration of funding within a limited number of institutions. Furthermore, involving external reviewers strengthens the alignment of national research evaluations with international standards of excellence. By embracing a more diverse and internationally oriented reviewer pool, agencies can enrich the evaluation process and ensure a comprehensive and fair assessment of the research outputs.

Recommendation 7: The scoring across the exercise needs to be normalised according to disciplinary practices. In the higher education system, some institutions focus on disciplines that traditionally show lower publications patterns and scarce use of metrics, which lead to a lower average scoring. Such institutions are therefore penalised when there are no normalisation processes applied to harmonise all disciplinary areas in the evaluation. This is a best practice that the VQR has already

implemented but that needs to be monitored in order to prevent any disciplinary biases.

USE OF AI TECHNOLOGY

Available commercial platforms offer search engines to match reviewers to papers under review. In particular, some Web of Science Reviewer Locator allows editors to quickly browse through a shortlist of possible reviewers, selected by the locator on the basis of their previous publications and reviews. According to the Clarivate website the algorithm implemented in the Locator trawls the extensive Web of Science dataset, including publications, citations, and peer reviews, to return up to 30 precise recommendations from over 28 million authors. Reviewers are located on the basis of their full publication history and potential organisational and co-author conflicts are flagged in the results of the search. This search engine supports the choice of the reviewers but still requires the editor to make the final selection from a recommended list. Specific details on the underlying technology are not openly available and need to be requested. It should nevertheless be considered that tools based on Web of Science for example retain the limitations of the database, which currently holds only about 30% of SSH publications.

From a technical standpoint, the simplest approach to matching reviewers is through reviewer and paper classification. However, this would require an active role of the reviewers to enter their area of expertise and of the authors to enter the classification of the paper. An automatic matching on the basis of the content of the paper and the past publications of the reviewers, can be achieved through Natural Language Processing (NLP). The first step would be to apply a Name Entity Recognition (NER) model to extract entities from the research output under evaluation. An entity is a word that has a relevant meaning in the context in which the entity is extracted. The same NER model would need to be applied off-line to the database of reviewers to extract entities from their publications. Existing generic Large Language Models (LLM), like BERT, and the more recent GPT4, can be used for this task. Available open source tools like SpaCy could also be a solution. These generic LLM need to be customised to be applicable to the specific context of interest. Once entities are extracted one can proceed with an automatic classification of the research output and of the reviewers. Note that a Machine Learning classifier can be trained to automatically associate research output to reviewers from the extracted entities once the reviewers are available in a database with their associated classification or directly from their associated entities.

The Panel understands that the implementation of AI tools is not a trivial task, and, although specific technologies like NLP or existing platforms like Clarivate reviewer locator can be used, the adoption of AI for reviewer selection would require some dedicated time and effort.

Recommendation 8: Over time, ANVUR should carefully explore the use of AI tools to facilitate the assignment of experts to research outputs. While recognizing the value of partial automation, it is essential that GEV panel chairs and vice-chairs retain the autonomy and opportunity to review such mechanisms.

Recommendation 9. The panel recommends that ANVUR explores the possible use of existing commercial platforms, in the first place. Alternatively, a dedicated platform based on Natural Language Processing (NLP) can be developed over time with the appropriate support of the Minister. Both options will require a dedicated effort and can be considered for the future.

USE OF REVIEWERS AND SCORING SYSTEM

The panel agreed that the use of peer review evaluation is the correct approach to assess the quality of the research outputs. The intention of this section is to highlight how the accuracy of the scoring of the research outputs is affected by the evaluation process. In this context an accurate evaluation of the outputs would consistently return a correct scoring of the outputs based on quality, where quality is measured by three criteria: rigour, impact and novelty.

As explained in the following section an evaluation process solely based on bibliometrics would return a consistent albeit possibly incorrect result. On the other hand, in the ideal case in which reviewers are not reliant on the bibliometrics, an evaluation based solely on peer reviews would be uncertain but possibly correct, if the reviewers operate without bias. An evaluation in which reviewers are affected by the bibliometrics would return an uncertain scoring with a lower uncertainty because the assessment is conditional on the value of the bibliometrics.

From a statistical point of view, the peer review process can be seen as an expert elicitation process, with the possible dependency on the metrics. This means that each output should be assessed by a large number of reviewers who should be diverse in gender, age and geographical distribution. Reviewers should work in isolation and a confidence interval should be computed on the totality of the scores returned by the reviewers. The classification of the outputs should then be based on the confidence interval and not on a simple consensus approach.

EVALUATION CRITERIA FOR DISCIPLINARY FIELDS

The evaluation criteria used in the last VQR are appropriate and are applicable to all disciplinary fields. However, their weight and interpretation does vary from discipline to discipline. Their quantitative evaluation is also not unique and obvious to communicate to the reviewers. Probably the most difficult to quantify and potentially the most variable across disciplines is the academic or scientific impact.

The originality might need to be better qualified to distinguish between what is often understood as incremental innovation compared to ground breaking research. Where the former is normally a translation of an existing concept into new applications, the latter is a new foundational methodology, idea or approach, sometimes with no immediate application. The degree of originality can also include aspects like discipline hopping in which one methodology born and used in one discipline is transformed to address problems in a different discipline. In other disciplines, the progression is not so linear, and originality can be interpreted in a variety of other models: apply methods of one discipline to another context; provide a new reading of an analysis with new evidence; reinterpret text and models which have been accepted for a long time for example. It is advisable that examples of originality are discussed by the panels as part of the briefing to better inform the reviewers and the panel on how to translate the degree of originality into a proper evaluation.

The panel acknowledges that assessing methodological rigor can be multidimensional, but it is relatively easier to evaluate in STEM disciplines. To ensure reproducibility and verifiability of results, the use of open science practices is highly recommended. It is also important to determine if claims are supported by results, new evidence, data, and research pathways. Other quantifiable methodological aspects include the presence of experimental and theoretical results, statistical relevance of data and results, new methodological approaches, and interpretations. Reviewers and panels should engage in a discussion about rigorous examples from different disciplines during the initial briefing.

Scientific impact, being the most challenging criterion to evaluate, can vary significantly across disciplines. Even when limited to the scientific community (as social and economic impact are assessed separately), predicting the timeframe for a particular result's impact is difficult. ANVUR could consider the following options:

a) Specifying the timeframe for an output's impact and requesting evidence of its potential or achieved impact, including other aspects of scientific impact. This approach mirrors the assessment of engineering outputs in REF but risks focusing on short-term impact.

b) Accepting the variability in interpreting the potential impact of research outputs, resulting in a wider range of scoring. (See further considerations on scoring and transcoding.).
c) Providing specific examples and guidelines based on disciplinary context and national priorities to support knowledge creation and development in different economic sectors.

In summary, it is crucial for GEVs members to operationalize the criteria within their disciplinary contexts and align them with national priorities for research and knowledge creation.

ASSIGNMENT OF SCORES AND TRANSCODING TABLE

In general, the panel has found the scoring table appropriate. However, the statistics on the scoring of the reviewers can be highly uncertain (see section on the number of reviewers and the uncertainty in expert elicitation).

The higher granularity of the scores and the number of reviewers per output can lead to a large uncertainty in the scores. The transcoding table offers a possible mitigation but the sole use of the median or average is not sufficient. Another problem is the weight of the criteria. An output with a "limited" rigour (which might suggest wrong results or unsupported claims) could still score well if originality is "excellent" and the perceived impact is also "very good". In this respect ANVUR might want to consider some minimum thresholds on rigour. The notes seem to suggest that thresholds on average scores per criterion are already applied but it is unclear how they are applied to the example above.

ANVUR might also consider the following. The 5 transcoded categories present a variable variance (or size of the confidence interval). For example, in the same 5 points the categories change from standard to excellent. This implies that reviewers correctly assess excellent outputs but fail at assessing non eligible outputs as "non-eligible" is not a category. Given the uncertainty on the quantification of impact and the possible variability in the scoring of originality a high accuracy down to a single point might be difficult to achieve.

ANVUR might want to consider proper thresholds on rigour and consider the uncertainty in the evaluation of the other two criteria when considering the transcoding.

USE OF BIBLIOMETRICS

The use of bibliometric indicators to inform the peer review process is often unavoidable. However, the appropriate selection and use of bibliometrics is very important and in line with the notion of responsible metrics as a way of framing appropriate uses of quantitative indicators in the assessment of research (see Introduction for context on responsible research assessment). There is a wide agreement that this needs to be regulated and standardised so that the scoring of the reviewers is not biased by the metrics and the metrics are not quantitatively affecting the scoring. As reflected by DORA some bibliometrics like the Journal Impact Factor are highly debatable and their use is not recommended. Journal level bibliometrics such as journal ranking, JIF, SJR and SNIP **are also not recommended** because of the following limitations:

a) Journal ranking, Impact Factor and SJR may be independent from the quality of the individual papers they publish and also may not directly correlate with the peer review quality.
b) Multidisciplinary journals have different rankings according to different disciplines. Even if the VQR assigned them to a specific ASJC on the basis of the citations, the room for error and inconsistency is still high.
c) Influencing scores and longevity cover a long-period and average citations across all the papers published in a given journal, thus they might not be applicable to individual articles, especially if recently published.
d) Although indicators like the SNIP are to be preferred to the JIF thanks to its proportionality factor which accounts for citation distribution density, and it remains a poor indicator for multidisciplinary journals.
e) Most indexes do not include all sources and do not differentiate between negative and positive citations.

Paper specific metrics also need to be considered with care for the following reasons:

a) Products submitted towards the end of the VQR period will likely have a small number of citations.
b) The absolute number of citations is discipline dependent and remains appropriate as that the VQR will continue to normalise according to the disciplines.
c) Field Weighted Citation Impact (FWCI) of a single output is to be preferred but depends on the classification of the output.
d) Top percentiles are calculated from the citation index of choice (for example the FWCI). Top percentiles computed with the FWCI are consistent indicators if the classification

is correct. Furthermore, platforms like SciVal[8] automatically exclude self-citations, except when using the FWCI to compute the top percentile.

e) Only a small number of SSH disciplines fully accept bibliometrics for research assessment and many databases include only a small portion of SSH publications. Yet a responsible use of metrics can be included to support GEV evaluations if they find it appropriate and useful.

The use of journal-level metrics remains highly problematic and is not regarded as best practice internationally, as indicated in both the principles of DORA and the new Coalition for the Advancing of the Research Assessment (CoARA)[9]. The CoARA principle states that "This [the assessment of research] requires basing assessment primarily on qualitative judgement, for which peer-review is central, supported by responsible use of quantitative indicators.[10]

Recommendation 10: Journal-level metrics should not be provided as part of the assessment in the VQR in the future. ANVUR could select other indicators of their choice and ensure that all panels use them with the same criteria.

Exceptionally, in some cases, for example new articles, there may be some value in using journal-level metrics as a proxy for how good the review process of the journal is. Some recent studies argue that the JIF is, in some cases, better than the number of citations at identifying high quality articles[11]. The JIF is a better indicator when the peer review process of the journal is accurate, and the citation count is inaccurate. This is a logical consequence of the JIF being a cumulative indicator over all published articles, which leads to the same conclusion as above that trusting the journal ranking implies an assumption on the quality of the peer review process and consistency of article quality within a journal[12]. A counter example comes from a different study that shows that the "reliability of published research works in several fields may be decreasing with increasing journal rank".[13]

The use of appropriately contextualised article-level metrics such as FWCI is suitable for some fields. However, the panel considered also the cases where the use of bibliometrics in

[8] https://www.elsevier.com/solutions/scival
[9] CoARA https://coara.eu
[10] https://coara.eu/agreement/the-agreement-full-text/
[11] Waltman L and Traag VA. *Use of the journal impact factor for assessing individual articles: Statistically flawed or not?* [version 2; peer review: 2 approved]. *F1000Research* 2021, 9:366 https://f1000research.com/articles/9-366.
[12] The peer review quality of a journal determines the threshold of entry to the journal, but there will be variation of quality among the articles in the journal. Data from the REF shows that the very highest JIF journals contain articles that are across the range of quality scores determined by peer review
[13] Brembs, B. (2018). "Prestigious Science Journals Struggle to Reach Even Average Reliability". Frontiers in Human Neuroscience. 12: 37. doi:10.3389/fnhum.2018.00037. PMC 5826185. PMID 29515380.

evaluation is less appropriate or could be highly disputed: interdisciplinarity and social sciences and humanities. Both areas share similar characteristics to a certain extent.

In interdisciplinary research, several new peer reviewed scientific journals are emerging and – more importantly – are developing around policy issues. There is a remarkable pressure from policy makers and science policy to science being strictly connected with our current and future societal challenges. Publications emerging from externally funded projects are often truly interdisciplinary efforts and they also are often encouraged to publish in open access. The result is that many interdisciplinary, multidisciplinary and cross disciplinary articles are by nature published in non-subject related journals and not always in the highly scored scientific outlets.

Interdisciplinary articles pose an additional complexity in terms of citation patterns. These articles have the potential to be cited by different research communities, and the citation frequency may heavily depend on the readership and the adoption of the journal's contents. This, in turn, impacts the calculation of normalised indexes, which are designed to assess the impact of papers within specific disciplines rather than the broader study area that benefits from the research. When interdisciplinary articles are cited, the citing community can vary based on the readership and the specific interests of different fields. This means that the same article may receive citations from multiple disciplines, reflecting the diverse range of communities that find value in its content. However, when normalised indexes are calculated, they typically assign papers to a specific disciplinary category for evaluation. As a result, the assessment is conducted against the closest disciplinary context, rather than considering the broader interdisciplinary nature of the research.

Consequently, the normalised indexes may not fully capture the true impact of interdisciplinary research. They may fail to reflect the wide range of communities and fields that benefit from and cite these articles. This discrepancy underscores the need for evaluation frameworks and metrics that can account for the nuances and complexities associated with interdisciplinary research.

In Social Sciences, Art and Humanities (SSAH), the use of metrics can deeply vary from one discipline to another. Economics, Business/Management or Psychology journals for example have a well-established use of metrics and citation-based metrics can be accepted as relevant. This is due to the relative homogeneity of the scholars and the school of thoughts which are well connected and in a close network with each other. The other fields within social science and humanities instead have a very wide number of journals with very different niche of specialisation in some cases and, as a result, their citation rates are scattered according to journal readerships.

The panel also identified that many of these fields and sub-disciplines use a wide range of languages and often are publishing in more regional and local peer reviewed journals given

the nature of their work. The study of society, social transformation and cultural heritage is often related to specific geographical areas and impacts specific regions and territories. Therefore, peer readership and citations are very well tied in with both disciplinary interests and logistical relevance.

In other words, the panel agrees that for the disciplines of SSAH, bibliometrics can be misleading in general, although in some disciplines retain some level of interest in research and especially researchers' evaluations.

Taking all the considerations above, great care is needed to encourage a truly responsible use of metrics and their interpretation of indicators to inform peer review. Yet recommendation 10 stands and metrics should not be provided but GEVs can be enabled to access metrics according to their needs and disciplinary practices and judgement to complement their assessments.

Recommendation 11: ANVUR should ensure that GEV members and external reviewers should receive specific instructions in case of use of bibliometrics and proper training to avoid the inappropriate use of some metrics.

The panel also considered whether there are other metrics, beyond those related to citation patterns, that could be useful in the assessment, specifically, so-called 'altmetrics'. Altmetrics generally relate to the mentioning of journal articles beyond academic spheres. Examples include mentions in Twitter posts, on blogs, in YouTube videos or in the media. While there is much interest in the use of altmetrics to determine the uptake of research in wider society, this is a relatively new area with not a precise meaning and validity for research assessment. The panel considers the use of altmetrics problematic and not appropriate.

Recommendation 12: ANVUR should not incorporate the use of altmetrics into the next VQR.

ASSESSMENT OF SOCIAL SCIENCES, ARTS AND HUMANITIES

The evaluation of SSAH disciplines shows some additional peculiarities compared to the evaluation in the STEM fields for a number of reasons. To those setting out to design a generally applicable research evaluation process, these SSAH disciplines remain challenging because their key characteristics are not easily 'quantifiable'. They include transfer of knowledge, or insights to society which are hard to measure. They have a greater diversity of dissemination practices and close interaction links to professions and society, making it hard to simply import approaches that have been successful elsewhere.

Many of the challenges raised within the SSAH community concern evaluation methods that are seen to be imported indiscriminately from other fields, with different traditions and deployed using methods and vocabulary that are not appropriate in SSAH disciplinary contexts. For example, only a small number of SSH disciplinary communities fully accept a bibliometric approach to assess research excellence (see also discussion of bibliometrics, above). Other metrics such as Journal Impact Factor, or h-Index, are rarely used. A few exceptions - despite pressure from science policy – are found among economists, or business/management scholars where such metrics-based approaches are widely accepted across researchers in the field and are more similar in character to some STEM fields. SSAH scholars also more often publish in non-English journals and in several languages which are underrepresented in many of the biblio-datasets underpinning metrics approaches. Multilingualism is a particularly positive aspect of some disciplines in this domain and considered as a sign of quality of the research or a path to open research results to specific communities. This was reflected in the ANVUR report which demonstrated that some of the SSAH domains publish in both English and Italian and other languages, far more frequently than STEM where English is more prevalent.

A pronounced diversity of methods, processes and applications even within the disciplines is what distinguishes research in the SSAH field. Also, research in several disciplines of this domain is not "linear" in the way knowledge is produced and accumulated, in other words the same nature of ever-changing societies means that the research revisits and returns on fundamental theory in view of profoundly changed conditions, social shocks and external factors. Finally, the local orientation of many SSAH disciplines has created some profound differences in research methods and applications across European countries.

Over the last 10 years, organisations like ENRESSH[14] have been monitoring and studying how to improve assessment of research across all disciplinary areas and how to develop new tools and methods which better evaluate the qualities of SSAH scholars. It is the assessment of the ENRESSH project that SSAH disciplines need adapted and transparent methods of evaluation to improve the understanding of how research in these disciplines generate knowledge, and the patterns of dissemination of research results in SSAH. The European funding programme Horizon Europe has already introduced some attention to social impact and public good of research but the implementation of SSAH integration in mainstream collaborative science programmes is still weak; mainly because evaluation approaches fail to capture the real contribution of such research. Furthermore, current evaluation approaches across other major EU programmes exclude meaningful participation and contribution of SSAH-driven ideas[15]. Currently, the overall evaluation of SSAH is in the area

[14] See https://enressh.eu/

[15] See EC monitoring reports of SSH integration showing that SSH integration in the challenges of Horizon 2020 is still rather weak European Commission, Directorate-General for Research and Innovation, Kania, K., Bucksch,

of interest of CoARA with a potentially dedicated working group addressing evaluation in these disciplines across the world. The outcomes would be instrumental for designing assessment procedures of SSAH in the next VQR.

Meanwhile, sustained attention must be paid to SSAH research practices and their evaluation as the disciplines themselves have changed following the impact of technology and big data applications. As mentioned, the SSAH disciplines are characterised by fundamental differences across schools of thought in different countries. As these disciplines rely heavily on peer review, it is legitimate to explore the question concerning the selection of reviewers and if a better alignment of different experts, especially when international scholars are involved, should be obtained. As discussed in the section on experts' databases and reviewers' management, the role of external reviewers is of critical importance. Furthermore, for these disciplines there are two aspects to be considered. International experts would provide some additional expertise (e.g., language being the most obvious); at the same time, as SSAH disciplines are more geographically different and more dependent on societal context, could also provide a very different view or evaluation of the research outputs. In fact, as mentioned, SSAH disciplines present some very strong research communities divided by school of thoughts, approaches, methods and so on and this heavily impacts on the assessment of a research product.

It is fundamentally true that the evaluation of research in these fields is a lot more driven by experts and peers than from the use of database and computational tools. Selection of panel experts is therefore the most relevant and difficult task. The ANVUR shows in their report a very high level of competence in the selection of a list of disciplinary experts, and a constant review of disciplinary panels. Yet, less reliable seems to be the random extraction of GEVs once the lists are compiled.

It is key – as suggested previously, that the selection of the panel members can be entrusted to the panel chairs and vice-chairs as those identified with a larger range of expertise in line with the coverage expected by a panel.

Recommendation 13: ANVUR should ensure that experts in research methods are included in the evaluation process either for briefing and/or within disciplinary GEVs.

As mentioned in the ANVUR report, research outputs, including publication practices, tend to be very different in different disciplines. In particular, SSAH disciplines are still publishing both in articles and monographs, although pressure from the science policy environment is pushing towards journal articles. The VQR still adopts a double scoring for a book versus an individual paper, acknowledging that the effort to execute a full monograph is definitely

R., Integration of social sciences and humanities in Horizon 2020 – Participants, budgets and disciplines, Publications Office, 2020, https://data.europa.eu/doi/10.2777/141795

greater. This is still recognised as best practices. In fact, inserting variable points for book size and type would require a further classification of publications which again means moving further away from international standards and focusing again more on metrics which are less reliable in these disciplines than in hard sciences.

In the ANVUR report it is suggested that data about the diffusion of the books in national and international qualified libraries are available, this could also be considered as a proxy for the quality of the output. However, there is a general agreement that there is no need to insert yet another quantitative method such as books distribution. One of the key arguments against the use of metrics is that they tend to measure popularity, rather than only rigour of research and quality of the output. Book distribution would be yet another element of the debate against use of the metrics as a support mechanism for peer review.

SSAH disciplines are also characterised by a wide variety of research outputs like for example musical compositions, art exhibitions, archaeological findings, different types of literature productions, and so on. There is no clear agreement about how such alternative outputs should be assessed in a national evaluation system. The debate though it is not just limited to SSAH disciplines but also to STEM who are revisiting the contribution of additional outputs to their research assessment. This is a work in progress. It would be recommended that for the next VQR it is entrusted to panel members to assess any additional research output submitted which from the results in the ANVUR report remains still rather limited at this stage. However, clear guidelines about a more holistic research evaluation which strongly encourages submission of diverse research outputs beyond publication may be recommended.

Recommendation 14: ANVUR should retain the double scoring of monographs compared to articles, regardless of their distribution. It should also encourage submission of different research outputs beyond publications through the guidelines but leave the assessment to the panel members' discretion.

One final note is about the importance of multilingualism particularly in these disciplines but also to be encouraged across all disciplines too. It is recommended that national research evaluations continue to raise awareness across all fields of science about the importance of multilingualism in practice of science both in scientific and academic communications as recommended by UNESCO. Guidelines and toolboxes should be provided to encourage and incentivise research carried out and communicated in all languages, and to address language biases in metrics, expert assessment and ranking. According to the CoARA principles, "changes in research assessment practices should enable recognition of the broad diversity of valuable contributions that researchers make to science and for the benefit of society … irrespective of the language in which they are communicated".

The ANVUR report highlights how publications in English seem to score higher than those in Italian or other languages. Such practice should be discouraged in the briefing of the evaluators and reviewers.

Recommendation 15: ANVUR should encourage a fair evaluation of research outputs regardless of the language in which they are produced reinforced by a statement about the neutrality of the evaluation in relation to the language of publication in the new guidelines.

OPEN ACCESS PRACTICES

The fraction of assessed products available in Open Access (OA) remains remarkably stable (53.6% - 54.3%, VQR Report, Fig. 2.3.11) throughout the evaluation period. Public Research Organizations score significantly higher (73.1%) than universities (51.4%) in terms of the proportion of OA outputs. If general increase of the popularity of OA publications is a national policy priority, then universities have to be encouraged to increase OA share in their outputs.

The general rule, that products financed for at least 50% with public funds must be made available in OA, has probably reached the limits of its impact. Still, among considered university products, approx. 21% are not OA because editors do not allow for OA distribution (VQR Report, Fig. 2.3.10). This group of outputs provides space for the increase of OA percentage (the small fraction that is not open due to an embargo is not significant, for all types of institutions, although it can vary significantly among disciplinary areas).

In the most recent REF exercise in the UK there was a requirement that some types of research outputs (journal articles and conference proceedings) were available OA in order to be submitted for assessment. The OA requirements included options for publisher-required embargo periods, and there was tolerance of a proportion of output (5%) to be non-compliant but still admissible. Nonetheless, evidence suggests that this approach had a transformative effect on the amount of OA content from UK authors[16]. In the context of the next VQR, it would mean the introduction of an upper limit for the fraction of non-OA products, in addition to the current requirements. It can also lead to shortening of embargo periods. The value of this limit should be carefully balanced: too low one could hamper publication of research results funded with external sources. Also, publication in OA usually costs more, so the possible funding gains coming from consequences of evaluation should at

[16] Chun-Kai (Karl) Huang, Cameron Neylon, Richard Hosking, Lucy Montgomery, Katie S Wilson, Alkim Ozaygen, Chloe Brookes-Kenworthy (2020) Meta-Research: Evaluating the impact of open access policies on research institutions eLife 9:e57067 (https://doi.org/10.7554/eLife.57067)

least compensate for these additional expenses. As the average non-OA products percentage is approximately 45% now, introduction of a limit of, for example, 40% at institutional level could provide a mild stimulus to increase the average fraction of OA among the products, and to shorten the embargo times.

Open Science involves sharing of data, programs, and so on. This should be a priority when data are generated with public funds. There are many available public repositories (i.e. Zenodo) where data can be deposited using the rules indicated in the FAIR principles (Findability, Accessibility, Interoperability, and Reusability)[17]. ANVUR should request that these "products" especially if submitted for evaluation should be made public under the FAIR rules and using repositories that comply with rules requested by Horizon Europe for example[18] .

Furthermore, a report published in March 2021, the Open Access Diamond Model study highlighted that out of the 29,000 OA Diamond journals worldwide (only 1/3 are in established indexes) 60% is in SSH, (22% in science, and 17% in medicine) and the majority of these journals (about 86%) publish less than 50 articles per year.[19] The report shows that SSAH disciplines are clearly oriented towards publishing in OA, yet the ANVUR report shows that such publications tend to receive a lower score on average from the reviewers. On the basis of national priorities on encouraging OA publications, some guidance may be provided to reviewers about different practices of publications and the wide landscape of journals publishing in OA.

Recommendation 16: ANVUR should consider introducing upper limits for the percentage of non-OA product and that data generated with public funds are made available in accessible repositories under the FAIR rules. Finally, as different disciplines have a different approach to OA publications, provide guidelines around assessment and scoring of OA publications.

THIRD MISSION

The so-called 'Third Mission' refers to the engagement of institutions and departments with a wider context to validate how not just research, but universities at large have societal impact including on national, regional, and local development. This concept is distinct from

[17] See doi 10.1038/sdata.2016.18
[18] See doi: 10.5281/zenodo.7728016
[19] Bosman, J., Frantsvåg, J.E., Kramer, B, Langlais, P.C, Proudman, V., (2021) *The OA Diamond Journals Study. Part 1: Findings* DOI:10.5281/zenodo.4558704, https://openresearch.community/documents/oadjs-findings

the academic or scientific impact of research (which is assessed in the VQR by the assessment of research outputs). Third Mission is also a broader notion than societal impact of research, which is assessed in some other national research assessments internationally (for example, the UK REF and the Engagement and Impact evaluation in Australia).

Following an exploration of evaluating Third Mission activities in the previous VQR, VQR 2015-19 included assessment of Third Mission activities the outcome of which influenced funding for the first time. The panel considers Third Mission an important element of the evaluation and commends ANVUR for its commitment to its inclusion, which is in line with international practice in research assessment. This is an important part of the VQR again well aligned with European policies on empowering higher education institutions to develop in line with the European Research Area, and in synergy with the European Education Area (Council Conclusions, 28 May 2021).

The approach to assessing Third Mission in VQR 2015-19 was based around the assessment of Third Mission case studies, an approach again in line with international practice. The case studies for the Third Mission impact needed to belong to one or more "fields of action". These are: a) intellectual and industrial property evaluation; b) academic entrepreneurship; c) technology transfer; d) production and management of artistic and cultural heritage; e) clinical experimentation and health protection; f) life-long learning and open education; g) public engagement activities; h) production of public goods and policy instruments for inclusion; i) innovative tools in support of open science; j) activities related to the 2030 UN Agenda for Sustainable Development Goals.

The case studies were assessed by a Third Mission panel, distinct from the panels used in the assessment of outputs. The Third Mission panel contained both experts from academia and those of broader society. ANVUR provided us with data that demonstrate reasonable levels of agreement between the review scores of different experts, with over 80% of case studies scoring no more than one scores apart. This suggests that the assessment was robust and reliable, although there was higher reviewer variance in some fields of action[20].

The panel notes that the Third Mission in VQR 2015-19 influenced only 5% of the funding allocated. This is reasonable given the novelty of including assessments of this activity. Given the importance of Third Mission, and the evidence supporting the robustness of the VQR assessment, there is scope to increase the funding allocated on the basis of the Third Mission in the future, if the policy objective is to enhance the impact that universities and research organisations have on society. While the proportion of funding allocated based on this element is a policy choice, internationally similar assessments are weighted between 15%

[20] The consistency between reviewers for different fields of action ranged from around 75% to 90% agreement within one score.

and 25%[21]. An increased funding allocation will likely encourage additional Third Mission activity, while also increasing the quality with which it is reported in case studies.

Recommendation 17: The panel recommends that consideration is given to allocate an increased proportion of funding on the basis of the assessment of Third Mission.

The assessment of Third Mission and societal impact from research is a hot topic internationally. As described, the approach taken by ANVUR, while being distinct and sensitive to the national context, is well aligned with international practice in this area. The panels therefore encourage continued alignment with international practices and definitions concerning societal impact. In particular, a report published in 2022 by the European Commission offers a clear set of indicators to monitor and evaluate key impact pathways[22]. There is also experience to draw on from the Engagement and Impact assessment carried out in Australia, the Hong Kong Research Assessment Exercise and the UK REF. Critical evaluation of practices and standards from elsewhere will be important, along with recognition of the specific Italian context.

Recommendation 18: The panel recommends that ANVUR continue to align their approach to Third Mission assessment with EU recommendations and practices in Australia, Hong Kong and UK where appropriate.

In VQR 2015-19, Third Mission case studies were submitted across all 10 of the fields of action described above. However, data provided to the panel by ANVUR show that the submission of case studies across the fields of action was not uniform, with the 'Public Engagement activities' category accounting for over 30% of the case studies[23]. And one field of action, 'Innovative tools to support open science' receiving only around 1% of the case studies. In this latter case it would be reasonable to reflect on whether to continue with this field of action, with open research perhaps being better reflected in other aspects of the assessment. Overall, the panel considered the fields of action to be reasonable and aligned with the broad definition of Third Mission.

A weakness of the approach of defining fields of action in advance is that it may exclude important and impactful Third Mission activities that do not fit into one of the categories or an intrinsic bias in the assessment of some categories considered by definition more important than others. We understand that the purpose of predefined fields of action is to

[21] The Hong Kong RAE 2020 weighted impact at 15%, Poland's research assessment system at 20%, and in the UK REF the weighting is 25% in REF 2021 (raised from 20% in REF 2014).

[22] European Commission, Directorate-General for Research and Innovation, Nixon, J., Study to support the monitoring and evaluation of the framework programme for research and innovation along key impact pathways: indicator methodology and metadata handbook, Nixon, J. (editor), Publications Office of the European Union, 2022, https://data.europa.eu/doi/10.2777/44653

[23] If case studies had been distributed equally between the fields of action each would have received 10% of the case studies.

ensure that assessors with appropriate expertise are available on the Third Mission panel. An alternative approach might be to require case studies to be classified using a larger number of predefined keywords by the ANVUR, which could give more flexibility and accommodate new or expanding areas of action, while still allowing case studies to be matched with assessors.

Recommendation 19: The panel recommends that ANVUR review the fields of action and possibly replace them with the use of predefined keywords for Third Mission case studies, and considers a more flexible approach to classifying case studies.

The panel noted that there was some variation in the scoring of case studies dependent on the fields of action. The 'Public Engagement activities' field of action received a score that was below the average, in contrast with fields of action relating to technology transfer, entrepreneurship, intellectual property and health benefits. There may be disciplinary differences in the contribution to fields of action (with SSAH disciplines more likely to be found in the public engagement category, for example). In the future it will be important to ensure that consistent standards are applied across all fields of action to reduce the risk of disciplinary bias.

Recommendation 20: The panel recommends that guidance is provided to ensure consistent scoring of Third Mission case study between the fields of action.

In VQR 2015-19 the Third Mission case study requirement was determined at the level of the whole institution. The number of case studies required depends on the number of departments, with universities requiring one case study for every two departments, and research institutes requiring one case study for each department. Data provided to the panel by ANVUR demonstrates that an effect of this approach is that the requirement for institutions of a similar size can be different. For example, for the largest universities the case study requirement varied between 13 and 30. Needing to produce a larger number of case studies is likely to put a downward pressure of scores and may introduce unfairness into the assessment. While the panel is strongly supportive of assessing Third Mission at the institutional level, we consider it important that the case study requirement is more directly linked to the size of institutions rather than their structure (i.e. number of departments). In the documentation provided, ANVUR proposes a number of possible approaches to achieve this outcome. In selecting the option, it will be important to balance having sufficient case studies for a broad and robust assessment with the additional effort required in producing more case studies.

Recommendation 21: The panel recommends that in future Third Mission assessments the number of case studies required should be directly linked to the size of institutions.

ANNEX: MEMBERSHIP OF THE EXPERT PANEL

Jean-Bernard Auby, Professor Law School, Science Po

Paola Bovolenta Director of the Centro de Biología Molecular Severo Ochoa, Consejo Superior de Investigaciones Cientificas (CSIC)-Universidad Autónoma de Madrid (UAM)

Sandra Di Rocco, Professor of Mathematics and the dean of the School of Engineering Sciences at KTH, Royal Institute of Technology, Stockholm (Sweden).

Steven Hill, Director of Research at Research England

Gabi Lombardo, Director of European Alliance for SSH

Marcin Palys, Professor of Chemistry, University of Warsaw and EUA Board Member

Massimiliano Vasile, Mechanical and Aerospace Engineering Strathclyde University