The new European reform of research assessment

Since being released in July 2022, an [Agreement on Reforming Research Assessment](#) has been signed by more than 400 European research performing and funding organizations. It is intended to guide a reform and mutual learning process within a coalition of its signatories, [CoARA](#). This policy brief analyses the agreement and provides recommendations for the next steps.

Gunnar Sivertsen (NIFU) and Alex Rushforth (CWTS)

1. The potential for implementation

The *Agreement on Reforming Research Assessment* (hereinafter: ARRA) addresses three contexts of evaluation:

1) Individual researchers as they apply for positions, promotions, or internal resources
2) Individual research proposals in applications for external funding
3) Research performing organizations and units.

ARRA was developed in collaboration between the European University Association, Science Europe, and the European Commission. Their involvement may strengthen the potential for implementation of the reform in two of the three contexts: While the members of the EUA are directly responsible for assessments in the first context, members of Science Europe are directly responsible for assessments in the second context.

The situation is different for the European Commission and its possible role in the third context. The assessment and funding of public research organisations is a responsibility within widely differing national research systems, as shown in overviews initiated by the Commission (e.g., Jonkers & Zacharewicz, 2016) and through mutual learning processes facilitated by the Commission (e.g., Debackere et al., 2018). These experiences and documents demonstrate that mutual learning is more feasible than shared European guidelines for organizational assessment and funding. The publications are not among the initiatives and literature listed in the European Commission's scoping report *[Towards a reform of the research assessment system](#)* (2021), which laid the basis for ARRA and mainly focuses on agreed reforms in contexts 1 and 2 above.

We reason that ARRA will be easier to promote (though of course not without challenges) in the contexts for assessment of individual researchers and individual research proposals. See section 7 below.

2. The core commitments and their agendas

By signing ARRA, the organizations are effectively committing to ensure that their research assessments will:

- recognize and reward the plurality of contributions researchers make to academic life (not just publishing and bringing in grant money)
- respect epistemic differences between research fields
- reward new (or newly emphasized) quality dimensions such as open science (broadly defined), research integrity, and societal relevance.

These commitments reflect two influential agendas in recent years. One of them is the agenda of the official European policy of *Open Research* as adapted to career assessment and development in research organizations. It is inspired by documents such as *[Evaluation of research careers fully acknowledging Open Science practices](#)* (European Commission, 2017), *[Research Assessment in the Transition to Open Science](#)* (European University Association, 2019), and national initiatives in the Netherlands, Finland, and Norway (Pölönen and Mustajoki, 2022).

The other influential agenda is expressed in the aim "to enable a move away from inappropriate uses of metrics". ARRA follows this agenda of responsible metrics by referencing the *[DORA declaration](#)* (2012), the *[Leiden Manifesto](#)* (Hicks et al., 2015), *[The Metric Tide](#)* report in the United Kingdom (Wilsdon et al., 2015), the Global Research Concil's [Statement on Principles on Peer Review/Merit Review](#) (2018), and the [Hong Kong Principles for assessing researchers](#) (Moher et al., 2020). With the organization of CoARA, ARRA might be more successful than the other initiatives in mainstreaming reform in the contexts of individual proposal and researcher assessments.

The two agendas were already becoming connected before ARRA. A recent study of responsible metrics as a professional reform movement (Rushforth &

Hammarfelt, 2022) shows that it evolved from mainly focusing on indicator use to a broader discussion of research assessment criteria and practices and the wider academic culture they help underpin. These movements promote a *re-legitimation* of evaluative bibliometrics, whereby they can still play a role *if used appropriately*. ARRA's relationship to bibliometrics is however ambivalent: while endorsing key texts and mantras from the responsible metrics movement (see Section 4, below), elsewhere the ARRA text is hostile towards bibliometrics.

3. An inappropriate field of research?

According to ARRA, "responsible use of quantitative indicators can support assessment where meaningful and relevant". Positive examples of such *indicators* are given, but none are publication-based. The term then changes from indicators to *metrics*, as in "journal- and publication-based metrics", with only negative examples. The third of four 'core commitments' in ARRA reads:

> *Abandon inappropriate uses in research assessment of journal- and publication-based metrics, in particular inappropriate uses of Journal Impact Factor (JIF) and h-index.*

ARRA warns that the use of such metrics "may negatively affect the quality and impact of research" and "result in a 'publish or perish' culture that falls short of recognising diverse approaches and could come at the expense of quality".

The term "journal- and publication-based metrics" comes close to a common definition of *bibliometrics*, a term never used in the document, which is a field of research where much professional work is invested in developing appropriate indicators for research assessment. Notably, the three recurring negative examples of metrics in ARRA (JIF, AIS, H-index) are not among indicators professionally developed and tested by the field. ARRA's sometimes hostile tone towards "journal- and publication-based metrics" risks tarring all forms of bibliometrics with the same brush as these discredited examples. To promote responsible development and use of *bibliometric indicators*, closer relations are needed between ARRA and the field of research that the *Leiden Manifesto* (Hicks et al., 2015) originated from: The annual [STI conference series](#) organized by the European Network of Indicator Designers. This builds upon ARRA's call for

development, monitoring and renewal of research assessment criteria, tools, and processes, to be informed by state-of-the-art "research on research" evidence (Commitment 10).

4. The reliance on peer review

The Leiden Manifesto defends the use of bibliometric information for research assessment in its first principle:

> *Quantitative evaluation should support qualitative, expert assessment. Quantitative metrics can challenge bias tendencies in peer review and facilitate deliberation.*

Much more optimistically, peer review is described in ARRA as "the most robust method known for assessing quality", while possible limitations are recognized in another optimistic sentence:

> *To address the biases and imperfections to which any method is prone, the research community re-assesses and improves peer review practices regularly.*

However, current problems with fatigue and distrust in peer review, as they might be experienced among the members of CoARA, are so far not addressed. How are the problems solved? What can the organizations learn from each other?

Also useful would be a review of the scientific literature of studies of under what conditions and with what possible constraints peer review works well in research assessment. Parts of this literature deals with the increasing problems with reviewer fatigue and distrust, and presents ideas about how they might be tackled. An overview pertaining to reviews in *external funding contexts* is given in Langfeldt (2021), while Reymert (2020) overviews the literature on criteria and practices in *academic recruitment*.

There is a tendency in ARRA to play quantitative versus qualitative off against one another. We suggest that the challenge is to find the best configuration of both in each assessment context.

5. Documentation for narratives

ARRA not only calls for qualitative rather than quantitative assessment. Another important hallmark is to broaden the basis for research assessment "beyond journal publications". A long list is provided of possible experiences, qualifications, and outputs to be recognized in a holistic assessment.

Earlier investigations of the options for including a wide range of qualifications and outputs in qualitive assessment have arrived at *narratives* provided by the applicant as the solution. ARRA will move in the same direction. The earlier studies observed *documentation* of the narratives as a challenge. ARRA will need to make the same observation and try to solve it.

ACUMEN, a large EU-funded project in 2011-14, addressed the problem with combining multiple qualitative and quantitative evidence sources for a broad assessment of the qualifications and outputs of individual researchers. The ideas behind the project were that narratives could not stand alone without documentation, and that the sources of evidence needed to be further developed and standardized to avoid large workloads in application and assessment processes. The team developed an "ACUMEN Portfolio" for the purpose and looked for relevant data sources. They found data from social media (as used in altmetrics) too limited in scope. They also investigated institutional research information systems as possible data sources without being able to implement the idea. A more recent project funded by Universities Norway, NOR-CAM, arrived at a similar possible solution, a flexible and interactive CV drawing on data from the Norwegian Current Research Information System, but so far without being able to implement it.

Our view is that CoARA could have an important mission in solving the problem with data sources for the documentation of broader qualifications and outputs.

6. Publications as documentation

ARRA gives outputs "beyond journal publications" much more attention than publications. There is a risk, therefore, that ARRA will disregard current developments in the scientific publishing system that may provide available information about research practices in a much broader sense than we are used to.

Publications are peer-reviewed and open to public discussion. The idea behind them is to make the research process behind the results transparent, open to criticism, and available for further use. The idea is often not followed in practice, but it can be reinforced. The developments in digital publishing allow for this.

ARRA lists items that should be assessed in addition to publications: data, software, models, methods, theories, algorithms, protocols, exhibitions. All of them are now publishable within a publication, in an appendix, or in linked documents. In fact, all *Indicators of responsible research practices* published with the Hong Kong Principles for assessing researchers (Moher et al., 2020) may now be represented in a scientific publication or by indicators derived from it.

ARRA also says: "Value a range of other contributions to responsible research and scholarly activity, such as peer review for grants and publications, mentoring, outreach, and knowledge exchange". Again, data sources and indicators for such activities are being developed within the scientific publishing system. Examples are those mentioned in the Annex of ARRA: Open science badges; Publons, ORCID, open peer review; CRediT; Reporting guidelines (e.g. EQUATOR Network) and metrics (Altmetrics, PlumX).

ARRA will need to clarify the value of scientific publications as documentation for research assessment. They demonstrate experience, achievements, and qualifications from performed research, and they may document many aspects of the research practices that ARRA will have problems with documenting from other information sources.

7. Differentiation of assessment contexts

ARRA rightly calls for differentiation between different aims and contexts of research assessment. The impression is nevertheless that the same main principles and commitments are applied in all contexts.

We find the guidelines in ARRA fully adequate for the assessment of *persons* as they apply for positions, promotions, or internal resources. Within research organizations, for their broad missions to

be fulfilled, there needs to be _Room for everyone's talent_, as is the title of a similar document preceding ARRA in the Netherlands.

The same ARRA guidelines may need adjustment and _concentration on research qualifications and their documentation_ to be adequate and practical for the assessment of _project proposals_ in contexts of research funding. The responsibility of funding organisations is to carefully select the most promising and innovative projects. They may also prioritize research themes independently of immediate institutional recruitment needs and in response to long-term societal needs. Some parts of a CV will be more relevant than other parts. Publications will be significant as documentation of experiences and achievements in performing research.

As indicated in the first section, ARRA tries to address organizational research assessment independently of the national systems in which such assessments are at work. Peer review and documentation serve other purposes in these contexts than they do in individual level assessments. Statistics (a term never used in ARRA) can be much more adequate here than in individual level assessments.

There is already an overload of summative organizational evaluations in the research sectors. Most of them are related to performance-based funding. Summative organizational evaluations look back at past performance, check whether goals or expectations have been reached, and serve decisions and/or resource allocation. Past performances are usually summed up from the individual to the organizational level. Formative evaluations, on the other hand, serve strategic development. They do not ask how individual researchers performed; they ask how the organization could improve in supporting good research (Sivertsen, 2023). ARRA is only focused on individual performances.

To further develop appropriate organizational research assessment, there is need for another document than ARRA. _CoARA_ could initiate such a document.

Policy implications

- There is need to develop a more constructive approach to bibliometric indicators.
- We suggest collaboration with researchers in the fields of research evaluation and indicator development.
- The increasing problems with reviewer fatigue and distrust need to be considered. Members of CoARA could create mutual learning about how they might be tackled.
- CoARA could have an important mission in solving the problem with data sources for the documentation of broader qualifications and outputs.
- There is need to clarify the value of scientific publications as documentation for research assessment. Current developments in the scientific publishing system may provide broader information about research practices.
- ARRA is adequate for the assessment of persons but needs adjustment for the assessment of research proposals. It is so far less adequate for research assessment at the organizational level.

Further reading

Debackere, K., Arnold, E., Sivertsen, G., Spaapen, J., Sturn, D. (2018). Mutual Learning Exercise: Performance-based funding of university research: Horizon 2020 Policy Support Facility, summary report. European Commission.

Hicks, D., Wouters, P., Waltman, L. et al. (2015). Bibliometrics: The Leiden Manifesto for research metrics. _Nature_ 520, 429–431.

Jonkers, K., and Zacharewicz, T. (2016). Research Performance Based Funding Systems: A Comparative Assessment. EUR 27837. Luxembourg: Publications Office of the European Union.

Langfeldt, L. (2021). The many dilemmas of grant peer review. R-Quest Policy Brief No. 6 (2021).

Moher, D., et al. (2020). The Hong Kong Principles for assessing researchers: Fostering research integrity. PLoS Biol 18(7): e3000737.

Pölönen, J., and Mustajoki, H. (2022).0'European recommendations on responsible research assessment. Presented at EASST Conference, Madrid, July 6-9, 2022. Session on Responsible Research Assessment and STS, organized by A. Rushforth, M. Sienkiewicz, J. Zuijderwijk, S. de Rijcke.

Reymert, I. (2020). Bibliometrics in academic recruitment: A screening tool rather than a game changer. Minerva, 1-26.

Rushforth, A., and Hammarfelt, B. (2022.). The rise of 'responsible metrics' as a professional reform movement: A collective action frames perspective. DOI: 10.31235/osf.io/cdmqz

Sivertsen, G. (2023). Performance-based research funding and its impacts on research organizations. In: Lepori, B., Jongbloed, B., Hicks, D. (Eds.) Handbook of Public Research Funding. Edward Elgar, Chapter 6, pp. 90-106.

Wilsdon, J. et al. (2015) The Metric Tide: The Independent Review of the Role of Metrics in Research Assessment and Management. doi:10.13140/RG.2.1.4929.1363.