https://doi.org/10.51408/issi2025_160

A Responsible Framework for an Appropriate Bibliometric-Based Research Assessment

Cinzia Daraio[1], Wolfgang Glänzel[2], Juan Gorraiz[3]

[1]*daraio@diag.uniroma1.it*
DIAG, Sapienza University of Rome (Italy)

[2]*wolfgang.glanzel@kuleuven.be*
ECOOM, KU Leuven (Belgium)

[3]*juan.gorraiz@univie.ac.at*
Dept Bibliometrics & Publication Strategies, University of Vienna (Austria)

Abstract

The growing reliance on bibliometric indicators in research evaluation has generated increasing criticism, both from the academic community and recent European initiatives advocating more holistic, peer review–centered approaches. This paper addresses the urgent need for responsible and contextualised use of such metrics. Rather than rejecting bibliometrics completely, we propose a conceptual framework that supports the appropriate application of bibliometric indicators, tailored to the goals, disciplinary contexts, and levels of analysis involved. This framework promotes a balanced approach, valuing transparency, interpretive care, and ethical use of quantitative indicators within broader evaluation systems. The paper, interpreting and substantiating CoARA (2022)'s claims, emphasises the integration of metrics with qualitative assessments to ensure academic integrity and societal relevance. It calls for shared protocols, cross-sector collaboration, and recognition of disciplinary diversity to ensure indicators *inform* rather than disappear from research assessment or dominate research assessment.

Introduction

In recent years, the application of quantitative approaches – particularly bibliometric indicators – in research assessment has come under intense scrutiny. Much of this criticism stems from concerns over the unintended consequences of these tools when used improperly. However, reform initiatives often lack conceptual clarity: they seldom define what exactly is being evaluated, at which level of aggregation, and with what granularity. This ambiguity leaves open whether "research" refers to a holistic academic process or merely to measurable outputs. Complicating matters further, much of the critique favoring peer review over metrics is based on issues observed at the individual researcher level – problems already acknowledged within the bibliometric community itself (Wouters et al., 2013).

This skepticism towards indicators has spurred a wave of manifestos and declarations—such as DORA and the Leiden Manifesto – advocating for more responsible and meaningful approaches to research assessment (Wilsdon et al. 2015; Biagioli and Lippman, 2020; Curry et al., 2020).

At the European policy level, calls for change have intensified. The European Commission's 2021 scoping report advocates for a re-evaluation of current systems and was foundational for the CoARA agreement in July 2022 (European

Commission, 2021; CoARA, 2022). While these initiatives mark significant progress, they do not offer concrete operational tools or criteria for responsible indicator use (see also Daraio and Maletta, 2025).

In response, this paper argues that bibliometric indicators should not be dismissed altogether. Rather, their "inappropriate" use – such as applying them in contexts for which they were never intended – should be the real target of reform (Glänzel, 2006). Bibliometric indicators are analytical tools developed through rigorous scientific methods within the fields of scientometrics and information science. Hence, discrediting them broadly is both unjustified and counterproductive.

What is required is a structured framework to determine whether the use of a specific indicator is fit-for-purpose in each evaluation context. The goal is not to oppose quantitative methods with qualitative ones, but to develop criteria that guide appropriate use, acknowledging that even peer review has limitations.

Thus, the paper proposes a multidimensional framework that outlines how indicators should be selected and applied responsibly in varying evaluation contexts. It concludes by identifying critical questions and limitations, while affirming the value of indicators – when used with expertise and care – in contemporary research evaluation.

Key Framework Dimensions for Evaluative Bibliometrics

In an influential contribution, Henk Moed (2017) introduced a visionary model of "evaluative informetrics," emphasizing how to practically apply bibliometric methods in research assessment. He later refined this framework, outlining the following four central questions essential to shaping evaluation studies.

1. What is the unit of assessment (e.g., individual, institution, country)?
2. What aspect of the research process is under consideration (e.g., scholarly impact, social benefit, interdisciplinarity, collaboration)?
3. What are the goals of the evaluation (e.g., resource allocation, performance improvement, strategic redirection)?
4. What are the characteristics of the assessed entities, including developmental stage or systemic relevance (Moed, 2020, p. 4)?

Table 1. The six dimensions of our Research Evaluation Framework.

#	Dimension	Definition	Warnings (or Pitfalls)
1	*Aggregation Level*	The scale at which evaluation is conducted: individual, group, institution, region, or nation.	Metrics must match the level: those valid at one level may mislead at other. Peer review suitability decreases with higher aggregation.
2	*Unit of Assessment*	The specific entity or profile being evaluated (e.g., individual researcher, lab, department).	Influenced by the context and nature of research; discipline and sector-specific needs matter.

3	*Purpose of Assessment*	The goal of the evaluation, such as funding, improvement, promotion, or benchmarking.	Drives methodology, timeline, baseline, and criteria. Different objectives call for different evaluation strategies.
4	*Context of Assessment*	The broader environment, conditions, and institutional or national environment in which research takes place.	Evaluations must be sensitive to systemic, geographical, or disciplinary contexts to avoid bias or misinterpretation.
5	*Elements of Research Process*	The stages and outputs of research, including input, process, output, and impact (academic and non-academic).	Must consider diverse impacts (e.g., social, economic, cultural) beyond scholarly output.
6	*Stakeholder Engagement*	Inclusion of those affected by or involved in research and evaluation: funders, institutions, public, etc.	Helps assess broader impact and legitimacy of evaluation; considers intended and unintended consequences.

Building on this foundation, we propose an expanded multidimensional framework by introducing two additional dimensions (see Daraio et al., 2024). Table 1 gives an overview of the proposed dimensions.

Criteria for Building and Using Research Evaluation Metrics Appropriately

The use of bibliometric and other quantitative indicators in research evaluation has grown increasingly complex. As shown in the Multidimensional Research Assessment Matrix (AUBR, 2010) and expanded by Moed (2017), there exists a broad array of indicators and methods intended to assess both scholarly and non-academic research impacts. While Moed (2017) offers concrete recommendations and evaluations of specific metrics, the AUBR matrix provides a more general overview of methods and their potential applications.

However, simply selecting from existing indicators is not enough. Even scientifically sound and well-designed metrics can lead to harmful conclusions if applied out of context. Therefore, the focus should not only be on building reliable metrics, but also on ensuring their appropriate application—tailored to the specific goals, level of aggregation, and nature of the research being evaluated.

To combine quantitative and qualitative methods meaningfully, diverse data types must be harmonized. Daraio and Glänzel (2016) proposed a standardized data integration model to support this process.

For bibliometric indicators to be meaningful and robust, they must meet several foundational conditions: i) data quality is essential; ii) metrics must ensure comparability (*commensurability*) and iii) results should be replicable over time (*validatability*). Bookstein (1997) further warns that measurement efforts are often undermined by randomness, ambiguity, and conceptual fuzziness. These challenges affect both metric design and interpretation.

To be considered fit for research assessment, indicators must meet a set of core criteria: they must be valid, meaningful, reliable, robust, and, where possible, normalisable and standardisable. This ensures that indicators are suitable for comparative evaluations and benchmarking.
Even after rigorous design, indicators must be applied within a conceptual framework that accounts for:

- The unit of analysis (e.g., researcher, institution),
- Disciplinary differences,
- Data infrastructure and publication behaviour.

Importantly, metrics must be selected based on their "fitness for purpose" – their ability to align with the specific assessment goals. Users should be aware of the margins of error they are willing to tolerate and interpret results in light of possible limitations or methodological flaws.
While both ex-ante and ex-post assessments are valuable, they require different types of data and interpretation. Therefore, a thoughtful balance between qualitative and quantitative approaches is essential. Qualitative aspects – like recognition, diversity, and societal engagement – must not be overlooked.
Finally, caution is advised when using composite indicators, which often suffer from non-transparency, arbitrary weighting, and component interdependence. Their tendency to compress multidimensional realities into a single value may obscure more than it reveals.
Table 2 offers a concise yet comprehensive overview of key dimensions that must be considered to ensure responsible, meaningful, and context-sensitive use of bibliometric indicators. It emphasizes that indicators should not be applied in isolation, but rather aligned with the purpose, unit of assessment, disciplinary norms, and stakeholder perspectives. By explicitly addressing methodological, interpretive, and ethical concerns – such as data quality, transparency, and fitness for purpose – the table supports evaluators in navigating complex assessment environments. It could be useful as a *practical checklist* or *diagnostic tool* to guide the informed and balanced application of metrics within broader evaluation frameworks.

Table 2. Criteria for the appropriate use of indicators in research evaluation.

Criteria	Key Elements/ Insights	Sources
1. Foundational Frameworks	- AUBR Matrix (2010) outlines multi-dimensional methods for assessing research performance. - Moed's evaluative informetrics (2017) provides practical applications, distinguishing academic and non-academic impacts. - Extends to alternative metrics for broader impacts.	AUBR (2010); Moed (2017)

2. *Appropriate Use*	- Indicators must be contextually appropriate – not all are fit for all settings. - Even valid metrics can mislead or harm when used improperly. - Importance of selecting metrics aligned with evaluation goals, level of aggregation, and disciplinary context.	General argument from paper; Glänzel (2006); Gorraiz et al. (2020)
3. *Data Integration*	- Combining qualitative and quantitative approaches requires harmonizing different types of data. - Standardized integration model proposed by Daraio & Glänzel (2016) to support coherent use of data in multi-purpose assessments.	Daraio & Glänzel (2016)
4. *Basic Data Requirements*	- *Quality***:** Data must be accurate, verified, and trustworthy. - *Commensurability*: Enables comparisons across cases, institutions, or disciplines. - *Validatability*: Results must be reproducible under identical data collection conditions.	Daraio & Glänzel (2016); Bookstein (1997)
5. *Measurement Pitfalls*	- *Randomness*: Unpredictable variability in measurement. - *Fuzziness*: Lack of clear definition or conceptual sharpness. - *Ambiguity*: Interpretational uncertainty. These issues affect both metric design and interpretive clarity.	Bookstein (1997)
6. *Indicator Criteria*	Indicators should be: - *Valid* – Measures what it claims to measure. - *Meaningful* – Yields interpretable, relevant insights. - *Reliable* – Statistically stable and reproducible. - *Robust* – Insensitive to minor changes in the system. - *Normalisable* – Adaptable to different scales. - *Standardisable* – Comparable and replicable across contexts. - *Quality-based* – Depends on high-quality data sources.	Moed (2017); Bookstein (1997); Daraio & Glänzel (2016); Gorraiz et al. (2016)
7. *Application Considerations*	Indicators must be aligned with - The *unit of assessment* (individual, institution, etc.) - The discipline's characteristics (e.g., citation practices) The *purpose of the evaluation* (e.g., funding, promotion) - Available infrastructure, data, and evaluation goals.	Moed (2017); EU Scoping Report (2021)

Table 2 (*contd*.). Criteria for the appropriate use of indicators in research evaluation.

Criteria	Key Elements/ Insights	Sources
8. Composite Indicators Warning	Should be used with caution**:** - Tend to obscure complexity. - May rely on *arbitrary weightings* and *inconsistent metrics*. - Risk loss of transparency**,** misinterpretation, and over-simplification. Interdependence of components may introduce systemic bias or noise.	General critique from paper; Moed (2017)
9. Balancing Methods	Responsible evaluation requires combining metrics with: - Peer review and expert input - Narratives and case-based evidence - Qualitative factors like diversity, recognition, and societal impact. Ensures fairness, inclusivity, and relevance across varied contexts.	Moed (2007); Best practice in research evaluation literature
10. Responsible use of indicators in research assessment	- Indicators must be applied with awareness of their limitations, context-dependence, and potential unintended consequences. - Requires critical reflection on indicator selection, data quality, purpose alignment, and fairness. - Must avoid mechanistic or symbolic use of metrics (e.g., compliance without reform). - Emphasizes transparency, reproducibility, stakeholder engagement, and ethical responsibility in interpretation and application. - Encourages use of indicators as *decision-support* tools, not decision-makers.	CoARA (2022); EU (2021); Moed (2017); Curry et al. (2020); General principles from the paper

An illustration of our framework

Figure 1 illustrates our framework that can be represented by an *optical prism*: The Prism of Research Evaluation. Just as a prism refracts white light into a spectrum of colours, the prism in this figure refracts the "light" of research performance through a structured and multi-dimensional evaluative lens. This figure signals a fundamental principle in responsible assessment: research quality is not a single colour or metric, but a multifaceted, context-sensitive construct. The three basic dimensions of our framework, the basis of our prism in Figure 1, from which to begin are: the unit to be evaluated (*whom we are assessing*), the research process to be evaluated considering its boundaries (*what we are assessing*), and the main goal of the assessment (*why we are doing the assessment*). We then have two important dimensions that allow us to specify *where*, *when*, and most importantly, *how* the assessment is carried out. They are the level of aggregation and the context of the evaluation, which constitute the two sides of our framework. Finally, we have the

dimension that completes our framework represented by all *stakeholders* interested in the evaluation and its impacts and effects (consequences). Our framework aims to apply some kind of spectral decomposition of the complex assessment task represented by light entering the prism. If it works correctly, the prism should provide a proper evaluation spectrum for the unit under assessment.

The refracted rainbow from the prism signals the diverse outcomes of evaluation when it is performed responsibly. No single metric or ranking can capture this plurality. Instead, we must strive to view research through multiple lenses, acknowledging that different purposes and contexts will yield different "colours" of insight.

This model embodies several key elements of responsible evaluation:

- *No one-size-fits-all*: Good assessment requires contextual fit between indicators and purpose.
- *Critical reflection*: Encourages evaluators to think through the boundaries and assumptions that structure assessment.
- *Participatory governance*: Promotes involvement of all stakeholders in defining meaningful metrics and methods.
- *Transparency*: Reveals how decisions are derived and reduces the black-boxing of evaluative procedures.
- *Indicator pluralism*: Supports a multidimensional approach to research assessment.

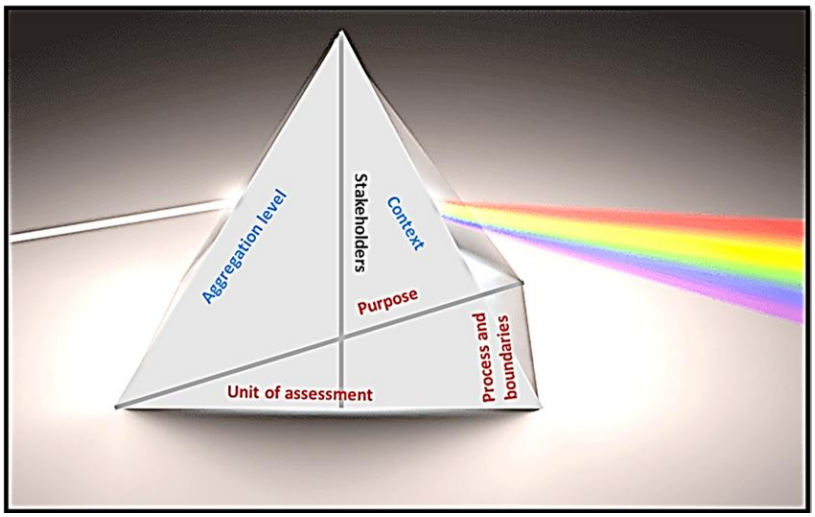

Figure 1. The Prism Model of Responsible Research Assessment.

Conclusions – Responsible and Contextual Use of Indicators in Research Evaluation

To ensure that bibliometric indicators are applied responsibly, a robust framework is essential – one that guides users in selecting the most appropriate metrics based on the specific goals, context, and evaluation problem at hand. The framework proposed in this paper aims to assist evaluators in choosing indicators that are fit for purpose

and in determining acceptable levels of uncertainty or error depending on the evaluation context. It also encourages the development of checklists to match available indicators to key assessment dimensions, thereby promoting structured and transparent decision-making (Robinson et al., 2024).

However, using the right indicators is not enough; they must be interpreted critically and carefully, with full awareness of the limitations imposed by methodological weaknesses, data quality, and parameter selection.

The shift from "publications and citations" to a broader spectrum of research contributions raises important concerns. What alternative outputs should be included in evaluation? How can we ensure these do not replicate the very problems that traditional citation-based metrics introduced – such as encouraging quantity over quality? For example, if researchers are evaluated based on uploaded outputs rather than impactful contributions, similar forms of metric manipulation could emerge. One proposal to counteract this issue, limiting the number of outputs submitted for evaluation, could help restore quality-based incentives. Yet such policies must be designed carefully to avoid unintended effects, such as disadvantaging early-career researchers or disciplines with rapid publication cycles.

Transparency and reproducibility must remain core principles in all evaluation methodologies. These can only be achieved if indicator use is standardized, well-documented, and paired with regular stakeholder interaction, including with researchers, institutions, and the wider community. Such engagement enhances both the meaningfulness and accuracy of the evaluation process and helps in identifying acceptable error thresholds and interpretive caveats.

To meet the complexity of today's research environment, bundles of valid and robust indicators should be selected, not created by arbitrarily combining metrics into opaque composite indicators. The paper cautions against composite indicators, as they often distort multi-dimensional realities, force linearity, and reduce transparency and interpretability. These effects directly conflict with the core principles of responsible metric use.

Recent approaches such as "narrative bibliometrics" (Torres-Salinas et al., 2024) offer a promising alternative. By embedding bibliometric data within contextualized, narrative interpretations, this method can enrich our understanding of impact, especially for less easily quantified outputs. Yet this, too, comes with limitations. The shift from objective metrics to subjective narratives introduces interpretive variability, which may undermine the neutrality typically associated with bibliometrics.

As Moed (2007) highlighted, the most effective evaluations combine "advanced metrics" with "transparent peer review". However, just as metrics require clear criteria for validity and reliability, qualitative evaluations also face challenges. Biases such as arbitrariness and fuzziness, critiqued by Bookstein (1997) in quantitative contexts, can also be present in peer review and narrative assessments.

Lastly, the growing role of Artificial Intelligence (AI) in bibliometrics introduces both opportunities and risks. AI tools can enhance data interpretation, detect meaningful patterns, and automate large-scale analyses. But they also risk reinforcing algorithmic biases, reducing human oversight, or narrowing the

evaluation lens. Any AI-based tools must be deployed with strong ethical guardrails, human interpretability, and accountability.

Rethinking research assessment is a complex but necessary undertaking. Incorporating a diversity of research outputs, improving the appropriateness of metric use, and embedding evaluation in ethical, transparent, and participatory practices are all critical. Achieving this will require not just methodological innovation, but active collaboration among researchers, institutions, funders, and policymakers.

Acknowledgement

This contribution is based on the conference paper by Daraio et al. (2024) presented as part of a special session organised by the authors at the STI 2024 Conference in Berlin. The objective of the present study is to contribute to the continuation of the debate within the framework of a special track at ISSI 2025 in Yerevan.

References

AUBR (2010), *Assessing Europe's University-Based Research – Expert Group on Assessment of University-Based Research*. (2010). Research Policy. European Commission. doi:10.2777/80193

Biagioli M, Lippman A eds. (2020), *Gaming the metrics: Misconduct and manipulation in academic research*, MIT Press.

Bookstein A. (1997), Informetric distributions. III. Ambiguity and randomness. *JASIS*, 48(1), 2–10.

CoARA (2022), *Coalition for Advancing Research Assessment*. (accessible at: https://coara.eu/app/uploads/2022/09/2022_07_19_rra_agreement_final.pdf)

Curry, S., de Rijcke, S., Hatch, A. et al. (2020), The changing role of funders in responsible research assessment: progress, obstacles and the way ahead. Working Paper. Research on Research Institute (RoRI) https://doi.org/10.6084/m9.figshare.13227914.v1

Daraio, C., Glänzel, W. (2016). Grand challenges in data integration—State of the art and future perspectives: An introduction. Scientometrics, 108(1), 391-400.

Daraio, C., Glänzel, W. (2020). Selected essays of Henk F. Moed. *Evaluative Informetrics: The Art of Metrics-Based Research Assessment: Festschrift in Honour of Henk F. Moed*, 15-67.

Daraio C., Gorraiz J., Glänzel W. (2024), Towards a framework for the appropriate use of bibliometric indicators in research evaluation, STI 2024 Conference, 18-20 September 2024, Berlin (Germany), https://doi.org/10.5281/zenodo.14036242.

Daraio C., Maletta S. (2025), Understanding Responsible Research Assessment: A MacIntyrean Proposal, *Acta Philosophica, International Journal of Philosophy*, 1, 34, 173-188.

EU (2021), "*Towards a reform of the research assessment system"*, EU Scoping Report. ISBN 978-92-76-43463-4. Accessible at: https://op.europa.eu/en/publication-detail/-/publication/36ebb96c-50c5-11ec-91ac-01aa75ed71a1/language-en).

Glänzel, W. (2006), *The perspective shift in bibliometrics and its consequences*. Accessible at https://de.slideshare.net/inscit2006/the-perspective-shift-in-bibliometrics-and-its-consequences

Glänzel, W., Debackere, K. (2003), *On the opportunities and limitations in using bibliometric indicators in a policy relevant context*, In: R. Ball (Ed.), Bibliometric

Analysis in Science and Research: Applications, Benefits and Limitations, Forschungszentrum Jülich (Germany), 225–236.
Glänzel, W., Schoepflin, U. (1994), Little scientometrics, big scientometrics ... and beyond? Scientometrics, 30(2–3), 375–384.
Gorraiz, J., Wieland, M., Ulrych, U., & Gumpenberger, C. (2020). De profundis: A decade of bibliometric services under scrutiny. *Evaluative informetrics: The art of metrics-based research assessment: Festschrift in honour of Henk F. Moed*, 233-260.
Gorraiz, J., Wieland, M., & Gumpenberger, C. (2016). Individual bibliometric assessment@ University of Vienna: from numbers to multidimensional profiles. arXiv preprint arXiv:1601.08049.
Moed, H. F. (2007). The future of research evaluation rests with an intelligent combination of advanced metrics and transparent peer review. *Science and Public Policy*, 34(8), 575-583.
Moed, H. F., Halevi, G. (2015). Multidimensional assessment of scholarly research impact. *Journal of the Association for Information Science and Technology*, 66(10), 1988-2002.
Moed, H. F. (2017). *Applied evaluative informetrics*. Berlin: Springer International Publishing.
Moed, H. F. (2020). Appropriate use of metrics in research assessment of autonomous academic institutions. *Scholarly assessment reports*, 2(1): 1. DOI: https://doi.org/10.29024/sar.8
Robinson-Garcia, N., Vargas-Quesada, B., Torres-Salinas, D., Chinchilla-Rodríguez, Z., & Gorraiz, J. (2024). Errors of measurement in scientometrics: classification schemes and document types in citation and publication rankings. *Scientometrics*, 1-21.
Torres-Salinas, D., Orduña-Malea, E., Delgado-Vázquez, Á., Gorraiz, J., & Arroyo-Machado, W. (2024). Foundations of Narrative Bibliometrics. *Journal of Informetrics*, 18(3), 101546.
Wilsdon, J., et al. (2015). The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management. DOI: https://doi.org/10.4135/9781473978782
Wouters, P., Glänzel, W., Gläser, J., Rafols, I. (2013). The dilemmas of performance indicators of individual researchers–An urgent debate in bibliometrics. ISSI Newsletter, 9(3), 48-53