

From Metrics to Reflexive Evaluation: Governing Science, Technology and Innovation Policy Without Losing Public Value

Hooman Shababi

Faculty Member, Rahedanesh Institute of Higher Education, Babol, Mazandaran, Iran

Hooman-shababi@rahedanesh.ac.ir

Submitted version (preprint) – not yet peer reviewed

This is a preprint published in fteval Preprints, the open repository of the Austrian Platform for Research and Technology Policy Evaluation (fteval). It has been screened for formal quality and thematic fit, but has not been peer reviewed. The content is the sole responsibility of the author(s).

Open review: This preprint has been submitted to the fteval Journal for Research and Technology Policy Evaluation and is undergoing open peer review. Review reports are published with their own DOIs and linked to this preprint. If the article is accepted, the final version will appear in the journal under a separate DOI and will be linked here. This preprint remains available in any case, together with its review reports.

How to cite this version:

Shababi, Hooman (2026) *From Metrics to Reflexive Evaluation: Governing Science, Technology and Innovation Policy Without Losing Public Value*. Working Paper. fteval Preprints. (Submitted)

Please check the DOI for the most recent version before citing: [10.22163/fteval.2026.768](https://doi.org/10.22163/fteval.2026.768)

License: CC BY 4.0

Abstract

Science, technology and innovation (STI) policy increasingly governs through indicators, rankings, benchmarks, evaluation rubrics and funding formulas. These instruments are useful because they make dispersed activities visible and support comparison, coordination and resource allocation. Yet they can also narrow policy attention, encourage strategic adaptation and displace the public, epistemic or social purposes that evaluation is intended to serve. This praxis-oriented article develops a framework for reflexive evaluation capacity. It argues that evaluation should not be treated as a neutral measurement stage at the end of a policy cycle, but as a socio-technical arrangement that actively shapes what institutions do and what policymakers can see. The article identifies five linked processes—valuation, commensuration, metricization, economization and performativity—and shows how they operate as a recursive valuation loop. It then proposes five design principles for policy evaluation: plural valuation, explicit theory of change, contestable indicators, protected spaces for qualitative evidence, and periodic review of metric effects. A practical implementation sequence is offered for ministries, funding agencies and research-performing organizations, illustrated with an empirical anchor drawn from Iran's evolving STI monitoring and evaluation architecture. The framework is positioned relative to the Leiden Manifesto, the Metric Tide, DORA and the CoARA Agreement, clarifying what it adds for STI policy actors working under resource constraints and in diverse institutional contexts. The aim is to help STI policy retain the coordinating benefits of metrics while reducing proxy displacement, gaming and the invisibility of public value.

Keywords: science and technology policy; innovation policy; research evaluation; metrics; valuation; public value; reflexive governance; responsible metrics

1. Introduction

Contemporary STI policy is increasingly conducted through calculative devices. Research systems are described through publication counts, citation indicators, rankings, grant success

rates, patents, commercialization measures, university-industry collaboration statistics and composite innovation indices (Hicks et al., 2015; Wilsdon et al., 2015). These devices are not simply descriptive. They help determine which activities receive attention, which organizations are considered successful and which forms of knowledge are treated as evidence (Espeland & Stevens, 2008; Miller, 2001; Porter, 1995).

The attraction of measurement is understandable. Policymakers must make decisions under conditions of uncertainty, limited budgets and competing demands. Indicators can make distributed activity more legible, support coordination across organizations and provide a common language for negotiation. They can also reveal inequalities, expose bottlenecks and help public institutions explain their choices (Porter, 1995).

The difficulty begins when a measure becomes a target, a funding condition or a reputational signal. Organizations then have reasons to adapt their behavior to the measure. The resulting improvement in the indicator may be interpreted as evidence that the underlying policy objective has been achieved, even when the relationship between the proxy and the objective is weak (Strathern, 2000; de Rijcke et al., 2016). The problem is not that indicators are always inaccurate. It is that they participate in producing the realities they claim merely to represent (Callon, 1998; Müller & de Rijcke, 2017).

This article develops a practical framework for reflexive evaluation in STI policy. Its central claim is that evaluation capacity should include the ability to examine the effects of measurement itself. An evaluation system should ask not only whether a target was reached, but also what kinds of work the target made visible, what it made less visible, how actors adapted, and whether the distribution of benefits remained aligned with public purposes (Bozeman & Sarewitz, 2011; Mazzucato, 2018).

The argument proceeds in six steps. Section 2 diagnoses the policy problem of proxy displacement. Section 3 elaborates a socio-technical framework built around five linked processes and shows how they form a recursive valuation loop. Section 4 positions the framework relative to the dominant responsible-metrics literature and specifies five design principles. Section 5 offers an implementation sequence and anchors it empirically in the Iranian STI monitoring and evaluation system. Section 6 discusses implications for policy capacity and reflects on limitations and costs. Section 7 concludes.

2. The Policy Problem: When Proxies Become Priorities

Evaluation systems usually begin with a legitimate concern: a broad policy goal must be translated into decisions. A ministry may seek to strengthen research quality, accelerate knowledge transfer, support regional development or address a societal challenge. Because these goals are heterogeneous and partly contested, policy actors select indicators that appear to provide usable evidence (Hicks et al., 2015).

The selection creates a chain of transformations. A multidimensional value is reduced to a set of observable properties. Those properties are standardized so that unlike activities can be compared. The resulting numbers are attached to decisions about rewards, sanctions or resource allocation. Organizations then invest effort in improving the measured properties. Over time, the proxy can acquire authority independent of the original policy purpose (Espeland & Stevens, 2008; Strathern, 2000).

This process is often described as gaming, but gaming is only one possible response. Actors may comply, reinterpret the measure, use it strategically, or develop workarounds that preserve

unmeasured objectives (de Rijcke et al., 2016). A university may improve its position in a ranking while reallocating effort away from locally valuable teaching or public engagement. A funding program may increase the number of collaborations while encouraging superficial partnerships. A national innovation strategy may raise patent counts while failing to improve the distribution of capabilities or the usefulness of technologies for underserved groups.

These are not arguments for abandoning measurement. They are arguments for recognizing that metrics are interventions. The relevant policy question is therefore not “Which single indicator is correct?” but “What configuration of evidence, judgment and institutional safeguards can support learning without allowing the proxy to replace the purpose?”

3. A Socio-Technical Framework for Evaluation

The framework proposed here connects five concepts that are often treated separately. The analytical contribution is not the invention of any single concept, but the demonstration of how they operate together as a recursive valuation loop that continuously reshapes STI policy attention and organizational behavior.

3.1 Valuation

Valuation is the attribution of worth. In STI policy, value may refer to scientific originality, economic productivity, social inclusion, environmental sustainability, national resilience or democratic accountability (Bozeman & Sarewitz, 2011; Mazzucato, 2018). These values do not always coincide. Evaluation becomes political because selecting one value for emphasis can redistribute recognition and resources. Making valuation explicit is therefore a precondition for any subsequent technical choice of indicators.

3.2 Commensuration

Commensuration makes different entities comparable through a shared scale or category (Espeland & Stevens, 2008). A publication, a community partnership and a policy contribution may serve different purposes, but an evaluation system may place them in the same scoring architecture. Commensuration can support transparency and coordination; it can also erase differences that matter for interpretation. The distinction from metricization is important: commensuration is the act of rendering unlike things comparable; metricization is the subsequent stabilization of that comparison into routine, quantifiable indicators.

3.3 Metricization

Metricization stabilizes selected properties as measurable indicators. It requires definitions, data infrastructures, classifications, reporting routines and thresholds (Miller, 2001; Porter, 1995). Once established, the metric can appear natural even though it reflects earlier choices about what counts, whose data are reliable and which time horizon is appropriate. Metricization therefore locks in the results of prior valuation and commensuration decisions, making them harder to reverse without deliberate institutional effort.

3.4 Economization

Economization organizes behavior and resources around calculative distinctions (Callon, 1998). It does not mean that every policy system becomes a market. Rather, actors begin to treat measured differences as grounds for allocation, competition, investment and strategic positioning. A ranking can therefore shape behavior even when no direct financial reward is

attached to it, because reputational and organizational consequences still flow from the measured differences.

3.5 Performativity

Performativity describes the feedback through which a measure helps produce the behavior it measures (Callon, 1998; Müller & de Rijcke, 2017). A metric is performative when the use of the metric changes the object of evaluation. This feedback may be beneficial: measurement can encourage coordination, reveal neglect or support collective learning. It may also be harmful when measured performance is improved at the expense of the broader purpose. Empirical work on “thinking with indicators” in the life sciences shows how performance metrics become integrated into the planning, organization and endpoint decisions of research itself (Müller & de Rijcke, 2017; de Rijcke et al., 2016).

3.6 The recursive valuation loop

Taken together, these five processes describe evaluation as an evolving arrangement rather than a final audit. The arrangement can be represented as a recursive valuation loop:

1. A policy actor identifies a problem and selects a value (valuation).
2. The value is translated into comparable categories (commensuration).
3. Selected properties are stabilized as measurable indicators with data infrastructures (metricization).
4. Institutions adapt practices and resources to the measured differences (economization).
5. New data are interpreted as evidence of performance; the interpretation strengthens, modifies or challenges the indicator’s authority (performativity), feeding back into renewed valuation.

The loop becomes risky when the gap between measured and valued outcomes grows while the evaluation system continues to treat the metric as authoritative. Reflexive evaluation capacity is the institutional ability to interrupt and revise this loop before proxy displacement becomes entrenched.

4. Positioning within the Responsible Metrics Debate and Five Design Principles

The five principles advanced below overlap with, but are not identical to, the major statements of the responsible-metrics movement. Clarifying the relationship is essential for a contribution aimed at STI policy practitioners.

4.1 Relation to DORA, the Leiden Manifesto, the Metric Tide and CoARA

The San Francisco Declaration on Research Assessment (DORA, 2012) focuses principally on the misuse of journal-level metrics, especially the Journal Impact Factor, in the assessment of individual researchers and articles. Its general recommendation—do not use journal-based metrics as a surrogate for quality—remains foundational, but its scope is narrower than system-level STI evaluation (DORA, 2012).

The Leiden Manifesto (Hicks et al., 2015) offers ten principles for the responsible use of quantitative indicators in research evaluation. It insists that quantitative evaluation should support qualitative expert assessment, that performance should be measured against the missions of the institution or researcher, that excellence in locally relevant research should be

protected, and that indicators should be scrutinized regularly. These principles are widely cited and form a common reference point for the fieval readership.

The Metric Tide (Wilsdon et al., 2015) introduced the language of “responsible metrics” organized around five dimensions: robustness, humility, transparency, diversity and reflexivity. Its UK-focused recommendations emphasized that metrics should support, not supplant, expert judgment, and that systemic effects of indicators should be anticipated and monitored.

The Coalition for Advancing Research Assessment (CoARA) Agreement (2022) is broader still: an institutional coalition whose ten commitments cover the whole research-assessment system—hiring, promotion, funding and organizational evaluation—and require signatories to publish action plans and report on progress. CoARA builds explicitly on DORA, Leiden and the Metric Tide while adding organizational accountability mechanisms (CoARA, 2022).

What this framework adds. First, it shifts the unit of analysis from research assessment (individual researchers, units, organizations) to STI policy evaluation as a governance practice that includes, but is not limited to, research assessment. Second, it supplies an explicit socio-technical account—the recursive valuation loop—of how metrics acquire authority and reshape behavior, drawing on valuation studies and the sociology of quantification rather than treating performativity as a residual risk. Third, it offers a concrete seven-step implementation sequence designed for ministries, funding agencies and research-performing organizations operating under resource constraints, including those in emerging and developing innovation systems. Fourth, it foregrounds public value (Bozeman & Sarewitz, 2011; Mazzucato, 2018) as the normative reference point against which metric effects should be audited, rather than treating “quality” or “excellence” as self-evident. The principles below therefore operationalize and extend the responsible-metrics agenda for STI policy settings where international rankings, composite indices and funding formulas already structure attention.

4.2 Five design principles

4.2.1 *Make valuation explicit (plural valuation)*

Every indicator system should state the values it is intended to support. A target for patenting, for example, may be justified by economic additionality, domestic capability, public health or private investment. These are different rationales with different implications. Making the rationale explicit allows stakeholders to contest it and reduces the tendency to treat a number as self-justifying (Bozeman & Sarewitz, 2011). An evaluation brief should therefore include a short valuation statement: whose benefit is being pursued, over what time horizon, and what trade-offs are considered acceptable. This statement should be revised when the policy context changes.

4.2.2 *Use a theory of change before selecting indicators*

Indicators should follow a theory of change rather than substitute for one. The theory should distinguish inputs, activities, outputs, outcomes and longer-term effects, while identifying assumptions and possible negative effects. It should also identify which parts of the causal chain cannot be captured by routine quantitative data. This approach prevents a familiar mistake: choosing indicators because they are available, then treating availability as evidence of relevance (Wilsdon et al., 2015). A small number of well-justified indicators, supplemented by qualitative evidence, is often more useful than a large dashboard with no clear decision logic.

4.2.3 Keep indicators contestable

Evaluation systems should create formal opportunities for evaluated actors and affected publics to challenge definitions, data sources and interpretation. Contestability is not an obstacle to evidence-based policy; it is a condition of learning because it exposes blind spots and makes disagreement visible (Hicks et al., 2015, principles 4–5). Contestability can be institutionalized through indicator review panels, public methodology notes, documented responses to objections and periodic revision windows. The purpose is not to give every claim equal weight, but to prevent the system from confusing administrative convenience with validity.

4.2.4 Protect qualitative and situated evidence

Some policy effects are relational, delayed or context-dependent. They may concern trust, institutional learning, inclusion, research culture or the usefulness of knowledge to a particular community. Such effects are difficult to compress into a single score. Evaluation systems should therefore reserve a defined role for interviews, narratives, peer judgment, case tracing and stakeholder testimony (CoARA, 2022, Commitment 2; Hicks et al., 2015, principle 1). Qualitative evidence should not be treated as an informal supplement that can be ignored when numerical results are available; it should have a declared function in the decision process.

4.2.5 Audit the effects of the metric

The final principle is to evaluate the evaluation system. At regular intervals, policy actors should ask whether the indicator is changing behavior, whether that change is desirable, and which outcomes are becoming invisible (Wilsdon et al., 2015, reflexivity; Hicks et al., 2015, principle 10). This review should look for metric displacement, strategic adaptation, unequal burdens, data exclusions and time-horizon distortions. The review should not wait for a crisis. An indicator can be retained while its use is modified, its weight reduced or its interpretation qualified. Reflexivity is therefore a continuing governance practice, not a one-time validation exercise.

5. An Implementation Sequence and an Empirical Anchor from Iran

The following sequence is designed for a ministry, funding agency or research-performing organization. It is deliberately modest in resource requirements and can be staged. To move beyond hypothetical illustration, the sequence is then applied to the evolving architecture of STI monitoring and evaluation in Iran.

5.1 The seven-step sequence

Step 1: Define the policy purpose. Write a one-page statement of the public, epistemic and institutional purposes of the program. Identify who benefits and who may bear costs.

Step 2: Map the valuation chain. For every proposed indicator, record the value it represents, the observable property used as a proxy, the data source, the decision attached to it and the behaviors it may encourage.

Step 3: Identify blind spots. Ask which activities, groups, forms of knowledge and longer-term effects are poorly represented. Treat missingness as a design issue rather than an absence of relevance.

Step 4: Establish a mixed evidence architecture. Combine indicators with peer judgment, document analysis, interviews, case tracing and feedback from affected stakeholders. Define in advance how each evidence type will influence the decision.

Step 5: Run a small-scale pilot. Before attaching strong rewards or sanctions, test the evaluation system in a limited setting. Observe how institutions interpret the measures and document adaptations that were not anticipated in the theory of change.

Step 6: Review metric effects. After each evaluation cycle, compare intended and observed behavior. Look for proxy displacement, data manipulation, exclusionary effects and improvements that do not survive outside the measurement context.

Step 7: Revise transparently. Publish a short account of what changed, why it changed and what evidence supported the revision. This makes the evaluation system itself a site of institutional learning.

5.2 Empirical anchor: STI monitoring and evaluation in Iran

Iran's STI governance system has undergone successive attempts to build national monitoring and evaluation capacity. Recent work associated with the Supreme Council of Science, Research and Technology (SCSRT) and the National Research Institute for Science Policy has moved from large indicator inventories toward more compact, theoretically structured frameworks. One third-generation framework, for example, reduced the indicator set from 112 to 40 indicators organized around an efficiency–effectiveness–utility model and incorporated a “beta testing” stage intended to support policy learning before full-scale deployment (Farazkish, 2025). Parallel efforts have sought comprehensive models that integrate enablers, governance and processes, outputs and impacts, with explicit weighting of governance quality and of research that addresses national challenges (Omidvar et al., 2026). Other contributions have proposed national frameworks for research excellence that give substantial weight to impact beyond the university, ethics and national-Islamic identity, and networked collaboration (IranDoc-related work, 2025).

These developments illustrate both the promise and the risks of metricization in an emerging innovation system. On the positive side, the reduction of indicator sets, the introduction of pilot or beta-testing stages, and the explicit inclusion of governance and societal-impact dimensions are consistent with Steps 3–5 of the sequence above. They also respond to a known problem: earlier generations of Iranian STI evaluation were often criticized for over-emphasizing scientific productivity and human-capital counts while under-representing innovation linkages, commercialization quality and problem-oriented impact (comparative analyses of Iran, Australia and New Zealand indicators, 2025).

At the same time, several tensions remain visible and map directly onto the recursive valuation loop. First, valuation remains partly implicit: composite performance scores and growth-rate rankings can still function as self-justifying targets even when the underlying public purposes (regional capability, health, environmental sustainability, inclusion) are only weakly specified. Second, commensuration and metricization continue to favor properties that are already countable—publications, patents, graduate numbers—over relational and delayed effects that matter for public value. Third, economization pressures are real: universities and research institutes operate under resource constraints and reputational incentives that reward measured improvement, even when that improvement is achieved by reallocation away from less visible work. Fourth, performativity is only partially audited: while beta testing is a valuable innovation, systematic post-cycle reviews of metric effects (Step 6) and transparent revision accounts (Step 7) are not yet institutionalized as routine governance practices.

Applying the reflexive framework therefore yields concrete recommendations for the Iranian (and similar) contexts. (1) Attach an explicit valuation statement to every major indicator set,

naming the public purposes and acceptable trade-offs. (2) Require a theory of change that distinguishes what the indicators can and cannot capture. (3) Protect qualitative and situated evidence—especially for problem-oriented research, regional missions and public engagement—by giving it a declared weight in funding and organizational decisions. (4) Institutionalize periodic metric-effect audits that examine gaming, displacement and exclusion, and publish short revision notes. These steps do not require abandoning quantitative monitoring; they require embedding it in a governance architecture that can learn from its own effects.

6. Implications for STI Policy Capacity, Costs and Limitations

Reflexive evaluation requires more than technical expertise in statistics or data management. It requires institutional capacity to connect measurement with policy judgment. Evaluation professionals need the ability to analyze indicators as socio-technical arrangements, while policymakers need the confidence to use evidence without demanding false precision (Porter, 1995; de Rijcke et al., 2016).

The framework also changes how accountability is understood. Accountability should not be reduced to proving that a numerical target was reached. It should include an account of the choices embedded in the indicator, the distribution of effects and the reasons for revising the policy instrument. This is especially important where STI systems are diverse, where national and regional priorities differ, or where policy objectives concern public value rather than immediate economic return (Bozeman & Sarewitz, 2011; Mazzucato, 2018).

For organizations in emerging and developing innovation systems, reflexive evaluation can help avoid importing indicators designed for different institutional contexts. International rankings and standardized benchmarks may provide useful orientation, but they should not be allowed to define capability or progress without contextual interpretation. A plural evaluation architecture can make local priorities visible while preserving opportunities for international comparison.

6.1 Costs, tensions and resource implications

Reflexive evaluation is not costless. The implementation sequence generates three principal tensions that practitioners must manage.

Evaluation effort versus available budgets. Mixed-evidence architectures, pilot stages and periodic metric-effect audits require staff time, data infrastructure and stakeholder engagement. In ministries and agencies with constrained evaluation budgets, these activities compete with routine monitoring. The practical response is staging: begin with valuation statements and valuation-chain maps (Steps 1–2), which are low-cost, before investing in pilots and full audits. Protecting a modest share of the evaluation budget specifically for reflexivity (for example, 10–15 percent) can prevent the reflexive components from being permanently deferred.

Speed versus learning. Policy cycles often demand rapid indicator production for annual reports or ranking exercises. Reflexive review slows the cycle. The tension is real; it can be mitigated by decoupling high-stakes allocation decisions from the most recent metric snapshot and by scheduling metric-effect reviews on a fixed, multi-year rhythm rather than continuous interruption.

Plurality versus comparability. Protecting qualitative and situated evidence improves validity for local purposes but reduces the ease of cross-organizational and international

comparison. The framework does not resolve this trade-off; it requires it to be made explicit and revisited. Composite indices can still be retained for orientation, provided their limitations are stated and their weights are open to contestation.

6.2 Limitations of the present contribution

Several limitations should be acknowledged. First, the Iranian empirical anchor draws on published framework documents and secondary analyses rather than primary fieldwork inside ministries or research-performing organizations; claims about organizational adaptation therefore remain indicative. Second, the five-process model is analytical rather than predictive: it organizes attention to metric effects but does not supply statistical tests of performativity. Third, the implementation sequence has not yet been subjected to controlled pilots by the author; its feasibility claims rest on alignment with existing responsible-metrics guidance and with design features already appearing in Iranian third-generation frameworks. Fourth, the treatment of public value remains programmatic: different polities will operationalize “public purposes” differently, and the framework does not prescribe a universal value hierarchy. Future work should test the sequence in one or more funding programs, document metric-effect audits in real time, and examine whether explicit valuation statements alter the distribution of resources across measured and unmeasured activities.

7. Conclusion

Metrics are indispensable to modern STI policy, but they are not innocent windows onto performance. They classify value, organize attention and shape behavior. Their benefits depend on the institutional arrangements in which they are embedded.

The practical alternative to metric fixation is not metric rejection. It is reflexive evaluation: explicit about values, grounded in a theory of change, open to contestation, capable of using qualitative evidence and willing to examine the effects of measurement itself. Positioned relative to DORA, the Leiden Manifesto, the Metric Tide and CoARA, the framework offered here supplies an integrated socio-technical account of how metrics acquire authority, a set of design principles tailored to STI policy governance, and a staged implementation sequence illustrated with the Iranian monitoring and evaluation experience. Such systems can retain the coordinating power of indicators while reducing the risk that proxies displace public, epistemic and social purposes.

The central test for an evaluation system is therefore simple: does it help institutions learn what matters, or does it merely reward what is easiest to count? STI policy that can answer this question openly is better positioned to govern innovation in the public interest.

References

- Bozeman, B., & Sarewitz, D. (2011). Public value mapping and science policy evaluation. *Minerva*, 49(1), 1–23.
- Callon, M. (1998). An essay on framing and overflowing: Economic externalities revisited by sociology. In M. Callon (Ed.), *The Laws of the Markets* (pp. 244–269). Blackwell.
- CoARA (Coalition for Advancing Research Assessment). (2022). *Agreement on Reforming Research Assessment*. <https://coara.eu>

- de Rijcke, S., Wouters, P. F., Rushforth, A. D., Franssen, T. P., & Hammarfelt, B. (2016). Evaluation practices and effects of indicator use—a literature review. *Research Evaluation*, 25(2), 161–169.
- DORA (San Francisco Declaration on Research Assessment). (2012). *San Francisco Declaration on Research Assessment*. <https://sfdora.org>
- Espeland, W. N., & Stevens, M. L. (2008). A sociology of quantification. *European Journal of Sociology*, 49(3), 313–343.
- Farazkish, M. (2025). Elaborating and empirical validation of the STI monitoring and evaluation framework for Iran. *Rahyaft*. <https://doi.org/10.22034/rahyaft.2026.11926.1583>
- Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, 520(7548), 429–431.
- Mazzucato, M. (2018). Mission-oriented innovation policies: Challenges and opportunities. *Industrial and Corporate Change*, 27(5), 803–815.
- Miller, P. (2001). Governing by numbers: Why calculative practices matter. *Social Research*, 68(2), 379–396.
- Müller, R., & de Rijcke, S. (2017). Thinking with indicators: Exploring the epistemic impacts of academic performance indicators in the life sciences. *Research Evaluation*, 26(3), 157–168.
- Omidvar, Z., Yaghoubi, N. M., & Kaskeh, S. (2026). A comprehensive model for evaluating the performance of the national research and technology system based on composite indicators. *Journal of Science and Technology Policy*. <https://doi.org/10.22034/jstp.2026.12344.2061>
- Porter, T. M. (1995). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press.
- Strathern, M. (Ed.). (2000). *Audit Cultures: Anthropological Studies in Accountability, Ethics and the Academy*. Routledge.
- Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Kain, R., Kerridge, S., Thelwall, M., Tinkler, J., Viney, I., Wouters, P., Hill, J., & Johnson, B. (2015). *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. HEFCE. <https://doi.org/10.13140/RG.2.1.4929.1363>

AI-use declaration

Generative AI was used under the author’s supervision for language drafting, structural assistance and literature organization. The conceptual argument, policy interpretation, empirical framing, references and final responsibility for the article belong to the author. The author reviewed and revised the manuscript before submission.