

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Submitted version (preprint) – not yet peer reviewed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 It has been screened for formal quality and thematic fit, but has not been peer reviewed.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. 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 organisations. The framework aims to help STI policy actors 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

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, commercialisation measures, university-industry collaboration statistics and composite innovation indices. These devices are not simply descriptive. They help determine which activities receive attention, which organisations are considered successful and which forms of knowledge are treated as evidence.

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 organisations and provide a common language for negotiation. They can also reveal inequalities, expose bottlenecks and help public institutions explain their choices.

The difficulty begins when a measure becomes a target, a funding condition or a reputational signal. Organisations then have reasons to adapt their behaviour 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. The problem is not that indicators are always inaccurate. It is that they participate in producing the realities they claim merely to represent.

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.

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.

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

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. A university may improve its position in a ranking while reallocating effort away from locally valuable teaching or public engagement. A funding programme 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 recognising that metrics are interventions. The relevant policy question is therefore not “Which single indicator is correct?” but “What configuration of evidence, judgement 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.

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. These values do not always coincide. Evaluation becomes political because selecting one value for emphasis can redistribute recognition and resources.

3.2 Commensuration

Commensuration makes different entities comparable through a shared scale or category. 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, yet it can also erase differences that matter for interpretation.

3.3 Metricization

Metricization stabilises selected properties as measurable indicators. It requires definitions, data infrastructures, classifications, reporting routines and thresholds. 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.

3.4 Economization

Economization organises behaviour and resources around calculative distinctions. 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 behaviour even when no direct financial reward is attached to it.

3.5 Performativity

Performativity describes the feedback through which a measure helps produce the behaviour it measures. 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.

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

- A policy actor identifies a problem and selects a value.
- The value is translated into a comparable indicator.
- Institutions adapt practices and resources to the indicator.
- New data are interpreted as evidence of performance.
- The interpretation strengthens, modifies or challenges the indicator's authority.

The loop becomes risky when the gap between measured and valued outcomes grows while the evaluation system continues to treat the metric as authoritative.

4. Five Principles for Reflexive Evaluation

4.1 Make valuation explicit

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.

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 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. A small number of well-justified indicators, supplemented by qualitative evidence, is often more useful than a large dashboard with no clear decision logic.

4.3 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.

Contestability can be institutionalised 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.

4.4 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 judgement, case tracing and stakeholder testimony.

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. For example, a funding review may use quantitative indicators to identify patterns and qualitative evidence to interpret mechanisms, distributional effects and unintended consequences.

4.5 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 behaviour, whether that change is desirable, and which outcomes are becoming invisible. 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

The following sequence is designed for a ministry, funding agency or research-performing organisation.

Step 1: Define the policy purpose. Write a one-page statement of the public, epistemic and institutional purposes of the programme. 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 behaviours 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 judgement, 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 behaviour. 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.

6. Implications for STI Policy Capacity

Reflexive evaluation requires more than technical expertise in statistics or data management. It requires institutional capacity to connect measurement with policy judgement. Evaluation professionals need the ability to analyse indicators as socio-technical arrangements, while policymakers need the confidence to use evidence without demanding false precision.

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.

For organisations in emerging and developing innovation systems, reflexive evaluation can help avoid importing indicators designed for different institutional contexts. International rankings and standardised 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.

7. Conclusion

Metrics are indispensable to modern STI policy, but they are not innocent windows onto performance. They classify value, organise attention and shape behaviour. 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. 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

- Callon, M. (1998). An essay on framing and overflowing: Economic externalities revisited by sociology. In M. Callon (Ed.), *The Laws of the Markets*. Blackwell.
- Espeland, W. N., & Stevens, M. L. (2008). A sociology of quantification. *European Journal of Sociology*, 49(3), 313–343.
- Miller, P. (2001). Governing by numbers: Why calculative practices matter. *Social Research*, 68(2), 379–396.
- 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.

AI-use declaration

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