2026, vol. 13, issue 1, pp. 74-82
RESEARCH ARTICLE https://doi.org/10.5281/zenodo.21036173

THE MEASURE IS THE ARENA: METRICS AS EPISTEMIC INFRASTRUCTURE

Dan Valeriu Voinea

University of Craiova

Abstract

Consequential metrics are commonly described as instruments for representing performance, quality, or progress. This article argues that, once embedded in institutional routines and linked to consequential decisions, metrics are better understood as epistemic infrastructures: durable arrangements of categories, proxies, data systems, interfaces, incentives, norms, and revision authorities through which value is made visible, comparable, and governable. The article asks under what conditions metrics shift from provisional proxies to infrastructures, and what epistemic and governance consequences follow from that shift. Drawing on infrastructure studies, the sociology of quantification, audit studies, global indicator research, and theories of performativity, the article develops a mid-range conceptual framework for this transition. It distinguishes metrics, metric regimes, and metric infrastructures; reconstructs a layered metric infrastructure stack; and identifies a capture process in which proxies, interfaces, incentives, adaptation, and institutional embedding become mutually reinforcing. Comparative discussion of machine-learning benchmarks, citation-based research assessment, and ESG ratings shows that metric infrastructures can coordinate inquiry and improve accountability while also producing proxy drift, false commensuration, epistemic narrowing, governance asymmetry, and lock-in. The article therefore treats metrics neither as neutral representations nor as intrinsically corrosive instruments, but as socio-technical arrangements whose epistemic value depends on how they are designed, governed, revised, and contested.

Keywords: quantification, metrics, epistemic infrastructure, audit culture, Goodhart's law, research assessment, machine-learning benchmarks, ESG ratings

1. Introduction

Metrics often enter institutional life as aids to judgment. A benchmark score summarizes the performance of a machine-learning system; a citation count offers a signal of scholarly uptake; a clinical endpoint records whether an intervention produces an intended effect. ESG ratings translate contested judgments about corporate responsibility into market-facing signals, while organizational dashboards make uncertainty appear manageable by converting it into numbers that can be compared, monitored, and acted upon. In each case, a metric appears to represent an underlying value such as capability, impact, health, sustainability, or growth.

The difficulty is that consequential metrics rarely remain mere representations. Once they are attached to ranking, funding, promotion, investment, regulation, or status, they become part of the environment in which actors make decisions. Researchers select venues, topics, and collaborations in fields where citations and journal prestige matter. Machine-learning teams choose datasets, model architectures, evaluation strategies, and claims of novelty in relation to benchmark leaderboards. Firms organize disclosure and managerial attention around rating categories that may vary substantially across providers. The metric no longer simply records action; it helps organize the arena in which action takes place.

This article conceptualizes that shift as a movement from measurement to infrastructure. A metric becomes infrastructural when it is embedded in databases, dashboards, ranking systems, reporting templates, professional norms, funding practices, and organizational routines. At that point, the question of validity widens. Beyond whether the number accurately represents a construct, it extends to whether the institutional system built around the number changes the construct, redistributes attention, narrows the field of legitimate evidence, or makes alternative forms of value less visible.

The article asks: under what conditions do metrics shift from provisional proxies to epistemic infrastructures, and what epistemic and governance consequences follow from that shift? The central claim is that consequential metrics should be analyzed as socio-technical arrangements through which communities know, compare, reward, and govern. Metrics can increase accountability, coordinate dispersed actors, make domains inspectable, and support cumulative improvement. They can also redefine the values they were intended to represent. The relevant analytical question is therefore not whether metrics are good or bad in general, but when they support reliable collective knowledge and when they quietly convert complex values into narrow operational targets.

The article makes three contributions. First, it distinguishes metrics, metric regimes, and metric infrastructures, thereby extending the discussion beyond proxy validity. Second, it develops a layered account of metric infrastructures and identifies a capture process through which categories, proxies, data systems, interfaces, incentives, adaptation, authority, and lock-in become linked. Third, it uses three analytically varied cases (machine-learning benchmarks, citation metrics, and ESG ratings) to show how the same infrastructural mechanism appears across technical research, academic evaluation, and market governance. The argument is conceptual rather than statistical: its purpose is to specify mechanisms, boundary conditions, and design principles that can orient future empirical work.

2. From Quantification to Infrastructure

Much of the critical literature on metrics begins with distortion. Goodhart's original monetary-policy formulation concerned the instability of observed relationships once policy begins to rely on them (Goodhart, 1975). Campbell's law generalized the problem for social indicators, arguing that the more an indicator is used for social decision-making, the more likely it is to be corrupted and to distort the process it is intended to monitor (Campbell, 1979). These accounts remain indispensable because they explain why actors adapt strategically when a measure becomes consequential. They also clarify why a metric that appears informative in a diagnostic setting may become unreliable when it is converted into a target.

Yet distortion does not exhaust the problem. Goodhart and Campbell explain what happens when a measure becomes an object of optimization. A theory of metric infrastructure must also explain how a measure prepares the ground for such optimization by defining categories, stabilizing comparisons, creating interfaces, and shaping the time horizons through which progress becomes recognizable. The point goes further: metrics help construct the very field in which gaming, compliance, competition, and legitimacy become intelligible.

The sociology of quantification provides the second foundation for this argument. Espeland and Stevens (1998, 2008) analyze commensuration as a social process through which qualitative differences are translated into common metrics. Such translation makes comparison possible, but it also changes social relations by deciding which differences matter and which can be ignored. Porter (1995) similarly shows how numbers acquire authority in settings where personal judgment is politically vulnerable: quantification can function as a technology of objectivity, especially when trust is scarce. Merry (2016) extends this analysis to global indicators, where numerical representations travel across contexts while stripping away local meaning; Davis et al. (2012) further show how indicators become technologies of global governance by influencing standards, decisions, and forms of contestation.

Audit studies add a closely related account of institutional verification. Power (1997) shows how audit extends beyond financial accounting into broader organizational life, becoming a ritual of assurance that reshapes the institutions it claims to inspect. Strathern (2000) demonstrates that accountability systems can transform internal values into externally inspectable forms, especially in academic and professional settings. Metric infrastructures overlap with audit cultures but are not identical to them. Audit studies emphasize verification, accountability, and assurance; the infrastructural account developed here emphasizes the broader arrangement through which measures become public arenas for comparison, competition, adaptation, and ontological closure.

Classification theory clarifies why metrics cannot be understood apart from the categories on which they rest. Bowker and Star (1999) show that classifications and standards help make the social worlds they appear only to describe. Before a score can be calculated, a field must decide what counts as a case, an observation, a success, a failure, or a residual category. These decisions are far from neutral, distributing visibility, embedding assumptions, and shaping whose work, experience, or harm becomes legible.

Performativity theory then shows how models and measures can become part of the worlds they describe. MacKenzie (2006) argues that financial models did not simply represent markets from the outside; under particular institutional conditions, they became incorporated into market practice. Metrics can operate similarly. A metric becomes performative when actors reorganize their activities around the conditions required to score well, so that the measure reshapes the object it is meant to represent.

Infrastructure studies supplies the final piece of the framework. Star and Ruhleder (1996) characterize infrastructure as embedded in practice, transparent in use, learned as part of membership, linked to conventions, and often noticed most clearly when it breaks down. Work on knowledge infrastructures likewise treats knowledge as produced through networks of people, artifacts, institutions, standards, and routines (Edwards et al., 2013). This perspective helps explain why consequential metrics become difficult to dislodge. Once a measure is built into databases, dashboards, reporting practices, funding procedures, rankings, and

professional identities, it no longer appears as a discretionary choice. It begins to appear as the practical reality of the domain.

The closest prior work to this argument is the literature on global indicators and the Sustainable Development Goals as epistemic infrastructures. Bandola-Gill et al. (2022) analyze the SDGs as epistemic infrastructures that connect numbers, networks, and governing paradigms. Tichenor et al. (2022) similarly use epistemic infrastructure to describe quantified global public policy. This article builds on that work by extending the infrastructural lens beyond global policy indicators and by specifying a mechanism through which metric regimes become competitive arenas across technical, academic, and market settings. The object of analysis is not the indicator alone, but the layered coupling of proxy, data system, interface, incentive, adaptation, revision authority, and institutional lock-in.

3. A Framework for Metric Infrastructures

A metric is a specific measure, such as benchmark accuracy, citation count, journal impact factor, h-index, graduation rate, hospital readmission rate, clinical endpoint, customer churn, net revenue retention, or ESG score. A metric regime is a set of related metrics and rules for using them: university rankings, machine-learning leaderboards, tenure evaluation procedures, clinical trial endpoint frameworks, startup dashboards, ESG rating methodologies, and policy scorecards are all examples. A metric infrastructure is broader still. It includes the metric and the regime, but also the categories that make measurement possible, the data pipelines that produce observations, the interfaces that display results, the institutions that rely on them, the incentives attached to them, the behavioral adaptations they elicit, and the authorities that maintain, revise, or retire the system.

In this sense, a metric infrastructure is a durable arrangement of categories, proxies, data systems, interfaces, norms, incentives, and revision authorities through which a community makes value visible and governable. The concept shifts attention from the accuracy of an isolated proxy to the institutional ecology that forms around it. A proxy with moderate validity may be useful when it is used locally, cautiously, and alongside other forms of judgment. A proxy with similar statistical properties may become epistemically damaging when it is publicly ranked, tied to high-stakes rewards, and treated as a surrogate for the value itself.

The difference is visible in research assessment. A citation count can serve as a descriptive signal among many forms of evidence. It becomes infrastructural when bibliometric databases, analytic platforms, rankings, evaluation committees, hiring decisions, grant reviews, and professional reputations begin to depend on it. At that stage, the count stops being a mere measure of scholarly influence and becomes part of the environment in which influence is pursued and recognized.

The layered structure of metric infrastructures can be represented as a stack. The stack is not a technical architecture in a narrow sense; it is an analytic device for identifying the levels at which measurement choices become institutional consequences.

Table 1. The metric infrastructure stack

Layer	Analytical question	Examples
Value	What is the underlying good or concern?	Health, intelligence, impact, sustainability, productivity
Construct	What concept is used to represent that value?	Clinical improvement, model capability, research quality
Proxy	What measurable indicator stands in for the construct?	Endpoint, benchmark score, citation count
Data	How are observations produced and stored?	Trial registry, leaderboard dataset, bibliometric database
Aggregation	How are observations combined or weighted?	Index, average, ranking, composite score
Interface	How are results displayed and compared?	Dashboard, leaderboard, profile, rating platform
Stakeholder	Who is measured, who measures, and who uses the result?	Researchers, firms, funders, raters, regulators
Incentive	What rewards or sanctions attach	Funding, promotion, status,

Layer	Analytical question	Examples
	to the number?	investment, compliance
Adaptation	How do actors change behavior in response?	Optimization, gaming, selection, reframing, disclosure management
Ontological	What comes to count as real progress?	State of the art, excellence, impact, responsibility, growth
Revision	Who can audit, contest, revise, or retire the measure?	Benchmark boards, journal policies, rating agencies, regulators

The stack shows why validity at the proxy layer is necessary but insufficient. A benchmark may validly measure performance on a particular dataset while still narrowing a field's understanding of capability if the leaderboard becomes the principal interface of recognition. A citation count may accurately count citations while still misrepresenting research quality if field differences, genre differences, language differences, or collaborative labor are collapsed into a single hierarchy. An ESG rating may faithfully implement a provider's methodology while still concealing that competing providers operationalize corporate responsibility in incompatible ways.

The central process can be described as metric capture. A complex value is translated into a construct, which a proxy then represents; the proxy is produced by a data system and displayed through an interface that makes comparison possible. Comparison attracts incentives, incentives produce adaptation, and adaptation changes the relation between proxy and construct. Institutions embed the proxy in routines, expectations, and identities, and over time the original value comes to be understood through the metric introduced to represent it.

This process is closely related to reactivity. Espeland and Sauder (2007) show that public measures can reshape the social worlds they claim to describe, especially when actors know they are being ranked. Dahler-Larsen (2014) similarly argues that performance indicators can have constitutive effects: they participate in forming the objects, expectations, and political relations later treated as measured facts. The infrastructural account gives those effects a layered explanation. It asks how the proxy, data system, interface, incentives, and revision arrangements jointly produce the conditions under which reactivity becomes durable.

Several theoretical expectations follow. Metrics become more constitutive as the target construct becomes more ambiguous. Research quality, model capability, sustainability, innovation, and excellence are all examples of values that cannot be reduced to one settled operational definition. Metrics also shift from representation to governance when they are tightly coupled to resource allocation. A privately used diagnostic indicator has different effects from a score used for hiring, promotion, investment, or regulation. Public comparability amplifies reactivity because actors adapt not only to improve an underlying practice but also to improve their relative position. Mature metric infrastructures produce metric-native practices: actors begin designing their work around what the metric will recognize. Finally, metric pluralism can reduce capture by preserving contestation, but it also increases coordination costs. A single dominant metric makes collective action easier; it also increases the risk that the metric becomes the field's operative definition of value.

This framework has boundary conditions. Not every metric becomes infrastructure. Local measures used for reflection may remain useful without becoming dominant. Measures that are weakly tied to rewards, frequently revised, openly contested, or deliberately subordinated to qualitative judgment may resist lock-in. The concept of metric infrastructure is most useful where a measure is durable, portable, embedded in institutional routines, linked to consequences, and used by multiple actors as a common basis for comparison or allocation. It is therefore a diagnostic concept: it asks how far a metric has traveled from measurement toward infrastructure, who maintains that infrastructure, who benefits from it, who must comply with it, and who has authority to revise it.

4. Comparative Cases: Benchmarks, Citations, and ESG Ratings

The following cases are selected for analytic variation rather than statistical representativeness. Machine-learning benchmarks are technical and scientific competition infrastructures; citation metrics are infrastructures of professional status and research assessment; ESG ratings are market and governance infrastructures built around contested moral and political values. The cases differ in their target constructs, users, incentives, and

forms of adaptation, but all show how metrics become embedded in public comparability, institutional reward, and behavioral response.

Machine-learning benchmarks

Machine-learning benchmarks illustrate the productive and narrowing capacities of metric infrastructure. A benchmark does more than evaluate a model. It defines a task, assembles a dataset, specifies an evaluation protocol, selects a score, establishes a comparison set, and displays an ordering of results. The leaderboard then becomes a public interface through which the field recognizes the state of the art.

This infrastructure is not merely coercive. Standard benchmarks allow researchers to compare results across laboratories, lower the cost of evaluation, and make cumulative progress more visible. Repositories such as Papers with Code organize papers, code links, datasets, methods, and evaluation tables in ways that increase the inspectability of a field (Papers with Code, n.d.). Benchmark repositories can therefore increase epistemic surface area: they make research activity easier to inspect, reproduce, debate, and extend.

The same arrangement can narrow attention. Raji et al. (2021) argue that broad AI benchmarks can come to stand in for ambitious and underspecified goals, making state-of-the-art performance appear to indicate progress toward general or flexible systems. Longjohn et al. (2024) similarly note that benchmark data repositories can improve benchmarking while also raising concerns about construct validity, representational harm, overreliance on a small set of datasets and metrics, and reproducibility. These concerns are infrastructural because they arise not only from individual datasets but from the system of recognition, comparison, and reward that forms around them.

The benchmark case traverses the full stack. The value layer concerns model capability; the construct layer operationalizes that value through task performance; the proxy layer is the benchmark score; the data layer is the dataset; the aggregation layer is the evaluation metric; and the interface layer is the leaderboard. Stakeholders include model developers, benchmark designers, dataset subjects, conference reviewers, firms, downstream users, and affected communities. Incentives flow through publication, status, investment, and institutional legitimacy. Adaptation appears as tuning, benchmark selection, data contamination, strategic framing, and the concentration of effort around measurable tasks.

Documentation practices such as datasheets for datasets and model cards respond directly to this infrastructural problem (Gebru et al., 2021; Mitchell et al., 2019). They require information about provenance, intended use, limitations, evaluation conditions, and subgroup performance, thereby reducing the burden placed on a single score. Such practices do not eliminate metric capture, but they make the infrastructure more inspectable and contestable. A healthier benchmark ecology would also require construct-validity review, benchmark retirement criteria, uncertainty reporting, stress tests beyond the main leaderboard score, and governance arrangements that include affected stakeholders as well as benchmark users.

Citation metrics and research assessment

Citation metrics show how measurement can become institutional legitimacy. Citations, h-indexes, journal impact factors, altmetrics, and rankings convert scholarly activity into portable signals. These signals can be useful in large and specialized research systems. They help institutions manage scale, reduce the opacity of expert judgment, and provide evidence of uptake. Their difficulty lies in what happens when they become tightly linked to hiring, promotion, funding, departmental status, and institutional prestige.

The San Francisco Declaration on Research Assessment was formulated in response to this problem. DORA states that the Journal Impact Factor was created to help librarians identify journals to purchase, not to assess the quality of individual articles or the contributions of researchers. It also identifies problems associated with skewed citation distributions, field specificity, manipulability, and the lack of transparent public data (DORA, 2013). The Leiden Manifesto similarly argues that quantitative indicators should support, rather than replace, expert judgment; that research should be assessed against institutional missions; that field variation must be respected; and that data and analysis should be open to scrutiny (Hicks et al., 2015). The Metric Tide report extends this position through the principles of robustness, humility, transparency, diversity, and reflexivity (Wilsdon et al., 2015).

Citation metrics have infrastructural power because they are portable and easily aggregated. Open bibliographic systems such as OpenAlex make works, authors, institutions, and other entities accessible through APIs and data downloads (OpenAlex, n.d.). Such infrastructures can increase transparency and reproducibility in bibliometric research. They also make citation-based comparison easier to routinize across committees, rankings, dashboards, and institutional reports.

The central epistemic difficulty is that citations are meaningful but heterogeneous. Bornmann and Daniel (2008) review studies of citing behavior and show that citations reflect diverse motives and social practices

rather than a single construct of quality. Citation practices vary by field, language, genre, publication culture, and collaboration pattern. Citations in mathematics, biomedicine, economics, and machine learning do not carry the same institutional meaning. Treating citation counts as a universal proxy therefore risks false commensuration: it makes different scholarly cultures appear directly comparable. De Rijcke et al. (2016) review evidence that indicator use in research evaluation can reshape academic behavior, including strategic responses and gaming. Biagioli and Lippman (2020) show that dependence on scholarly metrics can generate new forms of manipulation and misconduct.

The stakeholder structure is especially important. Researchers, journals, departments, and universities are measured; bibliometric databases, publishers, analytics firms, and ranking organizations produce or mediate the measures; hiring committees, grant panels, administrators, funders, and governments act on them. The residual category is large. Monographs, non-English scholarship, replication work, local public impact, slow scholarship, collaborative service, data curation, and mentorship often translate poorly into citation-based indicators. In a mature citation infrastructure, the practical question for many actors becomes not whether the metric is adequate, but how to improve their standing within it. This is the point at which a descriptive signal becomes a condition of professional practice.

ESG ratings

ESG ratings make the constitutive role of metrics especially visible because the object being measured is deeply contested. Environmental responsibility, labor practice, corporate governance, social impact, and supply-chain conduct are not naturally commensurable. ESG ratings nevertheless translate them into comparable and investable signals.

The empirical literature shows substantial divergence across providers. Berg et al. (2022) decompose ESG rating divergence into scope, measurement, and weight, finding that measurement explains the largest share of disagreement, followed by scope and then weight. Earlier work by Chatterji et al. (2016) similarly found limited convergence across social ratings, with implications for managers, investors, and researchers who treat different ratings as interchangeable. Christensen et al. (2022) further show that disagreement can persist around disclosure and interpretation. The deeper significance lies beneath the disagreement itself: rival providers often embody competing measurement ontologies, making different assumptions about what corporate responsibility includes, how it should be observed, and how its components should be weighted.

This case highlights the politics of metric infrastructure. When investors, index providers, firms, and regulators act on ESG scores, questions of category design, reporting burden, data quality, audit authority, and weighting become questions of governance. A rating does not simply report responsibility; it gives responsibility an operational form. If providers operationalize ESG differently while using the same general label, the label itself conceals a field of competing definitions.

The stakeholder asymmetry is pronounced. Rating agencies, data vendors, consultants, and disclosure platforms produce scores. Investors, index providers, regulators, and corporate managers use them. Firms, subsidiaries, workers, suppliers, local communities, and ecosystems are often the entities whose conditions are represented, aggregated, or omitted. Harms that are locally specific, difficult to quantify, embedded in supply chains, or experienced by communities without procedural power are especially vulnerable to disappearance within a composite score.

ESG ratings therefore illustrate the proposition that the more ambiguous the target construct, the more constitutive the metric becomes. The governance challenge is democratic and epistemic as much as technical: who defines responsibility, who determines weights, who can challenge data, who has standing to contest categories, and when a rating should be revised or retired.

5. Cross-case Analysis and Design Implications

Across the three cases, metric infrastructures differ in domain but share a common architecture. Each begins with a complex value: model capability, scholarly quality, or corporate responsibility. Each translates that value into operational constructs and proxies. Each uses databases, platforms, profiles, rankings, or dashboards to make comparison public. Each becomes consequential through its connection to rewards such as publication, status, promotion, investment, or reputational advantage. Each elicits adaptation. Each risks transforming the proxy into the practical definition of the value.

The cases also show that metric infrastructures are epistemically productive. Benchmarks make machine-learning systems comparable; citation databases make scholarly uptake visible; ESG ratings make corporate conduct legible to capital markets and policy actors. The problem is not visibility as such, but selective visibility that becomes difficult to contest once institutions build workflows and rewards around it.

Three cross-case findings refine the framework. First, construct ambiguity increases the constitutive force of a metric. ESG ratings are the clearest example, but the same issue appears in benchmark claims about capability and citation claims about quality. Second, interfaces format action. A leaderboard, bibliometric profile, or rating platform does not simply display information; it organizes attention, comparison, and aspiration. Third, revision authority is a central governance issue. A metric infrastructure becomes politically consequential when those who are measured, or those affected by measurement, have limited ability to correct data, challenge categories, contest weights, or demand retirement of the measure.

These findings suggest that reform should focus less on eliminating metrics than on governing the ecologies in which metrics operate. The relevant design question is how institutions can preserve the accountability and coordination benefits of measurement without collapsing complex values into single authoritative proxies.

Table 2. Pathologies and design responses in metric ecologies

Pathology	Institutional response	Tradeoff
Proxy drift	Review behavioral effects as well as statistical validity	Requires qualitative evidence and periodic reassessment
Goodharting	Loosen coupling between indicators and high-stakes rewards	Reduces administrative simplicity
False commensuration	Use plural indicators and require contextual interpretation	Makes comparison less efficient
Epistemic narrowing	Protect residual categories and unmeasured contributions	Slows evaluation and complicates decision-making
Metric lock-in	Establish sunset clauses, revision triggers, and retirement criteria	Creates uncertainty around continuity
Governance asymmetry	Give measured and affected actors correction, appeal, and participation rights	Can increase conflict and procedural cost
Overconfidence	Display missingness, uncertainty, caveats, and provider disagreement	Weakens the rhetorical force of rankings
Ontological closure	Publish construct definitions and revision histories	Prevents single-score finality

Three commitments follow. The first is epistemic humility. Metrics should travel as provisional stand-ins, not as definitions. They should disclose the construct they claim to represent, the assumptions built into their categories, and the dimensions they leave out. Qualitative judgment is not the opposite of accountability; in many domains it is what keeps accountability connected to context.

The second is institutional design against capture. Where values are genuinely complex, plural metrics preserve contestation better than a single score. This pluralism carries costs: it makes decisions slower, comparisons harder, and administrative systems less elegant. Those costs should be acknowledged rather than concealed. Coupling should also be managed. A measure used for learning should not automatically become a threshold for reward, sanction, or legitimacy.

The third is revision. If metrics are infrastructures, they need governance mechanisms comparable to those used for other infrastructures: maintenance, audit, versioning, appeal, and retirement. Validity should be assessed after a metric is deployed, not only at the moment of design, because the relation between proxy and construct changes once actors adapt. Rankings, scorecards, and dashboards should therefore make uncertainty, missing data, methodological limits, and disagreement visible. They should also specify who has authority to revise the system and under what conditions it can be retired.

6. Limitations and Conclusion

This article has developed a conceptual framework rather than testing a causal model. The cases were selected for analytic variation and are used to refine the argument, not to constitute a systematic sample. Future research could test the framework through comparative fieldwork, bibliometric analysis, benchmark histories, ESG rating changes, policy dashboard studies, or organizational case studies. It could also examine

negative cases: metrics that remain local, revisable, and non-infrastructural despite being useful. Such cases would help clarify the boundary between measurement and infrastructure.

Several issues require further development. The article does not provide a detailed account of law, regulation, litigation, or standard-setting around metric infrastructures. It also brackets many technical questions of measurement validity. Finally, it does not resolve how much metric pluralism is optimal when institutions must still make decisions. These limits are not peripheral. They indicate where an infrastructural account of metrics should be extended.

The argument is nevertheless straightforward. Metrics make progress legible, and that legibility is both their promise and their risk. They compress complexity, coordinate attention, support comparison, and enable governance at scale. When they become consequential, however, they also shape the worlds they claim to describe. They define the arena in which actors compete, the time horizon over which improvement is recognized, and the categories through which value becomes visible.

The concept of metric infrastructure captures this double character. Metrics are neither neutral representations nor merely corrupting targets. They are socio-technical arrangements through which communities know, compare, reward, and govern. Their power often lies in becoming ordinary: once embedded, the question shifts from whether the metric is appropriate to how performance on it can be improved. Governing measurement therefore requires more than better proxies. It requires attention to categories, data systems, interfaces, incentives, stakeholder rights, revision authority, and institutional lock-in. When the measure becomes the arena, designing the metric means designing part of the field's practical reality.

7. Declaration of Generative AI Use

During the preparation of this revised draft, the author used Claude (Opus 4.8, Anthropic) and OpenAI ChatGPT to assist with research-question refinement, structural revision, language editing, and citation cross-checking. These tools were not used to generate data, conduct empirical analysis, or originate the paper's scholarly claims. The author reviewed and edited the content and accepts full responsibility for the accuracy, originality, and integrity of the final manuscript.

8. References

Bandola-Gill, J., Grek, S., & Tichenor, M. (2022). Governing the Sustainable Development Goals: Quantification in global public policy. Palgrave Macmillan. https://doi.org/10.1007/978-3-031-03938-6

Berg, F., Koelbel, J. F., & Rigobon, R. (2022). Aggregate confusion: The divergence of ESG ratings. Review of Finance, 26(6), 1315-1344. https://doi.org/10.1093/rof/rfac033

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

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. Journal of Documentation, 64(1), 45-80. https://doi.org/10.1108/00220410810844150

Bowker, G. C., & Star, S. L. (1999). Sorting things out: Classification and its consequences. MIT Press.

Campbell, D. T. (1979). Assessing the impact of planned social change. Evaluation and Program Planning, 2(1), 67-90. https://doi.org/10.1016/0149-7189(79)90048-X

Chatterji, A. K., Durand, R., Levine, D. I., & Touboul, S. (2016). Do ratings of firms converge? Implications for managers, investors and strategy researchers. Strategic Management Journal, 37(8), 1597-1614. https://doi.org/10.1002/smj.2407

Christensen, D. M., Serafeim, G., & Sikochi, A. (2022). Why is corporate virtue in the eye of the beholder? The case of ESG ratings. The Accounting Review, 97(1), 147-175. https://doi.org/10.2308/TAR-2019-0506

Dahler-Larsen, P. (2014). Constitutive effects of performance indicators: Getting beyond unintended consequences. Public Management Review, 16(7), 969-986. https://doi.org/10.1080/14719037.2013.770058

Davis, K. E., Kingsbury, B., & Merry, S. E. (2012). Indicators as a technology of global governance. Law & Society Review, 46(1), 71-104. https://doi.org/10.1111/j.1540-5893.2012.00473.x

De Rijcke, S., Wouters, P. F., Rushforth, A., Franssen, T., & Hammarfelt, B. (2016). Evaluation practices and effects of indicator use: A literature review. Research Evaluation, 25(2), 161-169. https://doi.org/10.1093/reseval/rvv038

DORA. (2013). San Francisco Declaration on Research Assessment. https://sfdora.org/read/

Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., Burton, M., & Calvert, S. (2013). Knowledge infrastructures: Intellectual frameworks and research challenges. University of Michigan. https://deepblue.lib.umich.edu/items/7b6a177a-d533-4859-b9df-7a6c5d1d93eb

Espeland, W. N., & Sauder, M. (2007). Rankings and reactivity: How public measures recreate social worlds. American Journal of Sociology, 113(1), 1-40. https://doi.org/10.1086/517897

Espeland, W. N., & Stevens, M. L. (1998). Commensuration as a social process. Annual Review of Sociology, 24, 313-343. https://doi.org/10.1146/annurev.soc.24.1.313

Espeland, W. N., & Stevens, M. L. (2008). A sociology of quantification. European Journal of Sociology, 49(3), 401-436. https://doi.org/10.1017/S0003975609000150

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92. https://doi.org/10.1145/3458723

Goodhart, C. A. E. (1975). Problems of monetary management: The U.K. experience. In Papers in monetary economics (Vol. 1). Reserve Bank of Australia.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. Nature, 520(7548), 429-431. https://doi.org/10.1038/520429a

Longjohn, R., Kelly, M., Singh, S., & Smyth, P. (2024). Benchmark data repositories for better benchmarking. Advances in Neural Information Processing Systems, 37, 86435-86457. https://doi.org/10.52202/079017-2744

MacKenzie, D. (2006). An engine, not a camera: How financial models shape markets. MIT Press. https://doi.org/10.7551/mitpress/9780262134606.001.0001

Merry, S. E. (2016). The seductions of quantification: Measuring human rights, gender violence, and sex trafficking. University of Chicago Press.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220-229). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596

OpenAlex. (n.d.). OpenAlex developers: Overview. Retrieved June 28, 2026, from https://developers.openalex.org/

Papers with Code. (n.d.). paperswithcode-data. GitHub. Retrieved June 28, 2026, from https://github.com/paperswithcode/paperswithcode-data

Porter, T. M. (1995). Trust in numbers: The pursuit of objectivity in science and public life. Princeton University Press.

Power, M. (1997). The audit society: Rituals of verification. Oxford University Press.

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the everything in the whole wide world benchmark. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (Vol. 1). https://openreview.net/forum?id=j6NxpQbREA1

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. Information Systems Research, 7(1), 111-134. https://doi.org/10.1287/isre.7.1.111

Strathern, M. (Ed.). (2000). Audit cultures: Anthropological studies in accountability, ethics and the academy. Routledge.

Tichenor, M., Merry, S. E., Grek, S., & Bandola-Gill, J. (2022). Global public policy in a quantified world: Sustainable Development Goals as epistemic infrastructures. Policy and Society, 41(4), 431-444. https://doi.org/10.1093/polsoc/puac015

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://www.ukri.org/wp-content/uploads/2021/12/RE-151221-TheMetricTideFullReport2015.pdf