02/2026

DEval DISCUSSION PAPER

NATIONAL EVALUATION SYSTEMS UNDER REVIEW

Comparing Existing Assessment Tools and Introducing the National Evaluation Capacities Index (INCE)

2026

Sarah Klier

Juan Sanz

DEval
GERMAN INSTITUTE FOR DEVELOPMENT EVALUATION

IMPRINT

Published by

German Institute for Development Evaluation (DEval)
Fritz-Schäffer-Straße 26
53113 Bonn, Germany

Phonel: +49 (0)228 33 69 07-0
E-Mail: info@DEval.org
www.DEval.org

Authors

Sarah Klier[1]

Juan Sanz[2]

Bibliographical reference

Klier, S. and J. Sanz (2026), *National Evaluation Systems Under Review: Comparing Existing Assessment Tools and Introducing the National Evaluation Capacities Index (INCE)*, DEval Discussion Paper 2/2026, German Institute for Development Evaluation (DEval), Bonn.

ISBN 978-3-96126-257-1 (PDF)

The German Institute for Development Evaluation (DEval) is mandated by the German Federal Ministry for Economic Cooperation and Development (BMZ) to independently analyse and assess German development interventions.

DEval Discussion Papers present the results of the ongoing scientific study of evaluation and the effectiveness of development cooperation, thus contributing to relevant expert debates on evaluation, social science methods and development cooperation. The discussion papers are geared towards academics and practitioners in the field of evaluation, methodology research and development cooperation.

DEval Discussion Papers are written by DEval evaluators and external guest authors. In contrast to our evaluation reports, they do not contain any direct recommendations for German and international development organisations.

Although DEval Discussion Papers are internally peer-reviewed, the views expressed in them are only those of the authors and – unlike our evaluation reports – do not necessarily reflect those of DEval.

All DEval Discussion Papers can be downloaded as a PDF file from the DEval website:
https://www.deval.org/en/publications

[1] Team Leader for Evaluation Capacity Development at the German Institute for Development Evaluation (DEval)
[2] Evaluator at the German Institute for Development Evaluation (DEval)

Abstract

Systematic assessments of national evaluation systems (NESs) have been published at irregular intervals for almost 25 years. In an era in which the efficient use of evidence for political decisions is more important than ever, these assessments are very valuable since they help national governments and international development partners to identify areas for improving demand, supply and use of evaluation evidence in national decision-making processes. This discussion paper aims to systematically compare these works, including a new approach on measuring NESs – the National Evaluation Capacities Index (INCE) – on the basis of six criteria. INCE has been adopted in 11 countries across Latin America and is increasingly implemented in other regions across the globe. Compared to other assessments, INCE is implemented in a more participatory manner, and the instrument not only entails regular measurements of NESs' capacities but also encompasses a large peer-to-peer learning mechanism and allows for alignment and harmonisation among actors working to strengthen a given evaluation system through evaluation capacity development.

Keywords: Evaluation systems, evaluation capacity development, institutionalisation, evaluation culture, evidence-informed policy

CONTENTS

Figures

Tables

ABBREVIATIONS AND ACRONYMS

AI	Artificial intelligence
CLEAR-LAC	Center for Learning on Evaluation and Results for Latin America and the Caribbean (Centro para el Aprendizaje en Evaluación y Resultados de América Latina y el Caribe)
CONEVAL	National Council for the Evaluation of Social Development Policy (Consejo Nacional de Evaluación de la Política de Desarrollo Social, Mexico)
DEval	German Institute for Development Evaluation (Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit, DEval)
ECD	evaluation capacity development
EU	European Union
GEI	Global Evaluation Initiative
IDB	Inter-American Development Bank
INCE	National Evaluation Capacities Index
M&E	monitoring and evaluation
MESA	Monitoring and Evaluation Systems Analysis
MIDEPLAN	Ministry of National Planning and Economic Policy (Ministerio de Planificación Nacional y Política Económica, Costa Rica)
NES	national evaluation systems
OECD	Organisation for Economic Co-operation and Development
UN	United Nations
VOPE	voluntary organisation for professional evaluation
WFP	World Food Programme

1. INTRODUCTION: CONTEMPORARY CHALLENGES OF EVIDENCE-INFORMED POLICY FORMULATION

The use of evidence in political decisions has diverse challenges. In an increasing number of areas, evidence is often available in abundance, which makes it difficult to use efficiently. In other areas, there are serious evidence gaps that cannot be closed easily, particularly within tight time frames.

Non-specialist audiences frequently lack access to evidence, as complex research findings are rarely translated into digestible formats that facilitate rapid comprehension and application. This challenge is compounded by the inherent complexity of contemporary policy problems, which resist oversimplification without compromising analytical rigor. A further and growing concern is the challenge of identifying high-quality evidence among an array of information sources whose reliability is increasingly difficult to verify.

Last but not least, contemporary politics is witnessing the re-emergence of populism, which is openly hostile vis-à-vis research and evidence-based policy making, claiming that intuition leads to decisions that are superior to those based on technocratic evidence. This phenomenon represents a fundamental challenge to evidence-informed policy paradigms and threatens the institutional foundations of research–policy interfaces.

Against this backdrop, researchers have tried to strengthen the relevance of academic outputs for policy decisions, not only in publications but also through different initiatives and projects aiming to translate research into policy action. There are many initiatives to improve the production, synthesis and use of evidence, like the Evidence for Policy & Practice Information Centre (EPPI Centre) at University College London, the What Works Hub for Global Education at the University of Oxford, the Africa Evidence Network (AEN), the International Initiative for Impact Evaluation (3ie) and the Campbell Collaboration, to name only a few.

While numerous sources of evidence inform policy making, evaluation represents the only form of evidence that has been explicitly integrated into the policy cycle through institutionalisation. Consequently, the quality of evaluation systems constitutes a particularly relevant factor when examining evidence utilisation in political decision-making processes.

Evaluations are an important source of evidence for political decisions, which is why they have gained more influence over the last years. As Leeuw and Furubo (2008: 164) note: “Evaluation is a characteristic of modern states and, although it took some time to develop, it is now almost considered a ‘natural’ phenomenon.”

Over recent decades, diverse evaluation processes, structures and evaluation systems have emerged worldwide. This proliferation has been accompanied by a corresponding growth in research examining these processes, structures and systems. As part of this development, various assessment frameworks have been developed to systematically analyse and compare evaluation activities at national level.

Most recently, the National Evaluation Capacities Index (INCE) was introduced to assess evaluation activities, representing a significant departure from previously established instruments. This discussion paper aims to examine the utility of existing assessment tools, explore their distinctive features and analyse the key differences between traditional approaches and this new framework.

2. EVALUATION SYSTEMS AND THEIR CONTRIBUTION TO EVIDENCE-INFORMED POLICY MAKING

2.1 Evaluation culture, institutionalisation of evaluation and evaluation systems: A conceptual distinction

As policy evaluation activities have increased, so too has scholarly interest in systematically describing them and deriving implications for evaluation practice at national level. The resulting body of work addresses different research objects (see Figure 1), which emerged through a process of development and refinement over time.

Early studies examined all policy evaluation processes and structures within a country, including those that were formally institutionalised as well as those that had an informal nature or emerged from established practices. This research object can broadly be described as evaluation culture.

With the proliferation of formalised evaluation activities being embedded in the political-administrative system, interest grew in this very development – the institutionalisation of evaluation, or the embedding of evaluation activities in the policy cycle through formal structures and processes. Accordingly, the institutionalisation of evaluation is a second research object that has emerged.

It soon became evident that national evaluation systems (NESs) had developed in many countries, constituting yet another research object. The research object of NESs always encompasses the institutionalisation of evaluation, but also extends beyond it, taking into account further formalised structures and process outside the political-administrative system. The distinction between evaluation systems and evaluation culture is fluid, and the terms are not used uniformly.

Figure 1 Conceptual distinction between evaluation culture, institutionalisation of evaluation and NESs, based on the understanding of evaluation systems used in this discussion paper

Source: Deval, own visualisation

Leeuw and Furubo (2008) defined four criteria to distinguish evaluation systems from uncoordinated (albeit regular and high-quality) evaluation activity: (1) a distinctive epistemological perspective, (2) organisational responsibility, (3) permanence and (4) a focus on the intended use of evaluations.

While these four criteria remain valid, there is no universally accepted definition of evaluation systems. The following section outlines the understanding of evaluation systems adopted in this discussion paper.

2.2 What is an evaluation system?

An evaluation system is an institutionally bounded configuration of actors, organisations, rules (which may be implicit and explicit) and processes that systematically produces and absorbs evaluations to serve the functions of learning, accountability and knowledge generation.

This concise, brief definition of an evaluation system is helpful in illustrating what is meant by this term. However, this discussion paper aims to systematically compare assessments of evaluation systems that differ considerably from one another. Therefore, a more detailed description of evaluation systems is needed to explain the assessments' high degree of heterogeneity and also why evaluation systems are not as clearly delineated as the above definition might suggest.

For such a comparison of evaluation systems, two theoretical foundations appear particularly valuable: general systems theory and the agency–institution paradigm.

When describing an evaluation system, it is helpful to draw on the knowledge produced by systems theory. In general systems theory, a system is understood as a set of elements interacting with each other and, together, forming a whole, which means that a system is understood to be more than the sum of its parts. According to this theory, the different elements of a system (here, an evaluation system) interact with and influence each other. Relationships between system elements follow a certain pattern, and systems are open, in an exchange with their environment (Klier et al., 2022; von Bertalanffy, 1968).

The agency–institution paradigm describes the reciprocal relationship between acting agents (agency) and institutional structures (institutions). Institutions – understood as rules, norms and established practices – shape and constrain the actions of agents, but are simultaneously reproduced, interpreted and changed through the actions of agents. In an evaluation system, the interplay between agency and institutions gives rise to various tensions, of which only three are briefly presented here:

- **Quality vs. utility:** Institutional structures – such as standardised quality criteria and formalised processes – enable consistent evaluation quality, but may simultaneously constrain actors' capacity for context-specific adaptations to diverse user needs. However, actors may strategically navigate, resist or reinterpret these structures to enhance utility.
- **Production vs. absorption:** Institutionalised planning mechanisms – such as work plans and evaluation schedules – ensure predictable evaluation output, yet actors require flexibility to meaningfully appropriate, interpret and utilise evaluation findings within their specific contexts and according to their strategic interests. Rigid production schedules may conflict with actors' absorption capacities and timing needs.
- **Accountability vs. learning:** Formal accountability requirements and compliance mechanisms embedded in institutional structures often prioritise the reporting of demonstrable evidence over reflexive learning. While institutions emphasise standardised accountability functions, actors seek knowledge generation and adaptive learning, which require spaces for experimentation and dialogue as well as contextualised communication formats that institutional rigidity may inhibit.

These tensions are not simply technical trade-offs, but reflect the fundamental dynamic between institutional constraints and actor agency – institutions shape what actors can do, while actors continuously reproduce, adapt or challenge institutions through their strategic practices. An evaluation system must manage these tensions as optimally as possible.

Integrating the insights of both general systems theory and the agency–institution paradigm yields the following description of an evaluation system: An evaluation system comprises the dynamic interactions among diverse actors and organisations operating within (and sometimes outside) established rules, norms and processes of a given institutional configuration. System boundaries are typically defined by jurisdictional limits (for example, national, European Union [EU] or United Nations [UN] jurisdictions), though actors may strategically navigate or challenge these boundaries. Evaluation systems are shaped by explicit and implicit principles that reflect underlying power relations and competing institutional logics. They are also shaped by

(financial and human) resources available within the system and the competition for access to these resources. The latter is not limited to actors within the system as parallel systems may compete for the resources made available to the evaluation system stemming from the same central source. The result of resource allocation is often a reflection of underlying power relations within and between systems. The function of an evaluation system is to systematically produce and absorb high-quality evaluations, strengthening the core functions of evaluation – learning, accountability and the gaining of knowledge. These functions of evaluations may stand in contradiction to one another within an evaluation system. The rules and processes established within the evaluation system, including communication channels and opportunities for stakeholder participation, both enable and constrain the conduct of evaluations. Quality, relevance and utilisation are continuously negotiated among stakeholders with varying degrees of power and resources as well as different interests.

Based on this description, evaluation systems are characterised by inherent tensions and contradictions (quality vs. utility, production vs. absorption, accountability vs. learning and standardisation vs. contextualisation). These tensions are not merely technical challenges to be "balanced", but reflect underlying power relations and fundamental conflicts between different values, interests and institutional logics. A robust evaluation system must therefore provide spaces for negotiation, contestation and reflexive adaptation while also acknowledging that perfect equilibrium is neither achievable nor necessarily desirable.

This description shows that evaluation systems always exhibit a high degree of heterogeneity and that they constantly adapt. A systematic assessment of an evaluation system can therefore only ever be a snapshot, which is why repeated assessments are so valuable for learning about these systems.

2.3 The distinctive position of evaluation as an evidence source

It is important to emphasise that evaluations constitute only one of many sources for evidence-informed policy making. The diversity of available evidence is extensive, and the complexity of policy-making processes is considerable. Consequently, the expectation that policy making should be evidence-based – that is, relying on and strictly adhering to rigorous scientific evidence – is increasingly viewed as outdated. Contemporary scholarly discourse has largely shifted towards the more pragmatic concept of evidence-informed policy making, which considers scientific evidence as one important factor while simultaneously incorporating other elements, such as context, experience, values and practical considerations (Head, 2015).

However, as the definition of evaluation systems and their function makes clear, it is different for evaluation as a source of evidence. Evaluation is the only evidence source that is typically systematically institutionalised within the policy-making process, with the generated evidence being explicitly produced for this process. This gives rise to the claim that evaluation is an important part of evidence-based policy making, meaning that political decisions are actually grounded in evidence produced by evaluations rather than being informed by it.

2.4 External drivers and multilevel manifestations of evaluation systems

Evaluation systems manifest across multiple levels, including institutional, sectoral and national levels. Furthermore, there are actors that constitute important components of various evaluation systems but do not represent evaluation systems in their own right. These include bodies such as the Network on Development Evaluation (EvalNet) – established within the Organisation for Economic Co-operation and Development's (OECD's) Development Assistance Committee – which develops crucial guidance for many evaluation systems, and the United Nations Evaluation Group (UNEG), which aims, among other things, to strengthen the evaluation functions within UN organisations.

In numerous countries, the establishment of NESs has been externally initiated through the influence of international actors, norm diffusion or common external shocks. In particular, multilateral organisations, including the EU, development banks and UN agencies, have systematically conditioned their financial and technical assistance on the implementation of evaluation mechanisms to assess the utilisation and effectiveness of the support provided. In many cases, this conditionality approach served as a primary catalyst for the emergence of evaluation systems across multiple national contexts and has significant

implications for understanding the institutional origins, design characteristics and sustainability challenges of NESs in developing countries, where donor influence often plays an important role in shaping domestic policy and administrative structures (Feinstein, 2015; Jacob, 2023; Mackay, 2009; Meyer and Stockmann, 2020; Porter and Goldman, 2013).

2.5 Balance between evaluation institutionalisation and excessive bureaucracy

The institutionalisation of evaluation is a prerequisite for evaluation systems and presents both opportunities and challenges. While evaluation serves essential functions in promoting transparency, accountability and evidence-informed learning, sectors characterised by high evaluation intensity face the paradoxical risk that evaluation systems may become overly restrictive, thereby constraining evaluation's core functionality and adaptive capacity. This phenomenon was articulated by Dahler-Larsen and Raimondo in their opening address at the 2022 European Evaluation Society (EES) conference, where they highlighted the potential for highly institutionalised evaluation systems to become counterproductive for their intended purpose (see Dahler-Larsen und Raimondo, 2024). Their analysis suggests that excessive systematisation and bureaucratisation of evaluation processes may inadvertently limit evaluative thinking, reduce methodological flexibility and constrain the very learning and adaptation mechanisms that evaluation is designed to facilitate.

Also, the implementation of too many (strategic) evaluations poses a major disturbance to institutions. As shown by Klier et al. (2022), institutions need sufficient time to absorb an evaluation ("relaxation time"). Otherwise, paradoxically, conducting evaluations can pose risks to an evaluation system: too many evaluations can overwhelm an institution's capacity to absorb them, or the burden of the evaluation processes (including responding to and implementing the recommendations) can become so great that it leads to evaluation fatigue or even a degree of paralysis within the institution.

This tension between evaluation institutionalisation and functional effectiveness represents a critical consideration for organisations, sectors and national bureaucracies seeking to optimise their evaluation systems while maintaining evaluative rigor and utility. This phenomenon is typical of bureaucracies and was discussed early on in academic literature: Max Weber (1922) warns of the "inherent logic" of bureaucratic apparatuses, which quickly become difficult to control, and Robert K. Merton (1940) draws attention to the dysfunctional consequences of bureaucracies, such as rule compliance becoming more important than the original objective ("displacement of goals"), mechanical rule application occurring without regard to purpose ("ritualism") and specialisation hindering flexible problem-solving ("trained incapacity"). The risk of developing overly restrictive evaluation systems similarly exists at national level. NESs may suffer from over-bureaucratisation and, at the same time, face the challenge of implementation gaps, where evaluation regulations and policies exist formally but lack effective operationalisation and enforcement mechanisms, as Stockmann et al. detail in the Evaluation Globe's first volume (Stockmann et al., 2020a see chapter on Germany: 167-198).

The described area of tension (which is also described in subsection 2.2) underscores the importance of thoughtful evaluation system design that maximises the benefits of systematic evaluations while also mitigating the risks associated with over-institutionalisation or excessive bureaucratisation of evaluation processes at the institutional level and inadequate implementation at the national level.

Achieving an appropriate and flexible form of institutionalisation is therefore crucial for evaluation systems. However, the conceptualisation of evaluation systems underlying this analysis emphasises that institutionalisation represents only one dimension of an evaluation system, albeit a highly significant one. Additional dimensions of the evaluation system, such as the professionalisation of evaluation, the quality of evaluations, the presence of multi-agent spaces and the utilisation of evaluation findings, are equally determinative of the system and can, within a well-designed system, serve as safeguards against overly restrictive institutionalisation.

3. ASSESSMENTS OF NESs

3.1 Comparative examination of five diagnostic studies

In recent years, there has been a notable expansion of NESs, and this has been accompanied by scholarly and policy interest in their development and implementation (Pattyn and Bouterse, 2020). This trend reflects broader societal shifts towards evidence-informed governance, as articulated by Stockmann et al. (2022b: 1): "It is the climate of data-driven policy and an evidence-oriented society, why evaluation moved from sporadic application to solid institutionalisation in a variety of country contexts." While there has undoubtedly been a trend towards greater evidence utilisation, the assertion of an "evidence-oriented society" is overstated. In recent times, this development has been superseded by another phenomenon – evidence being portrayed as falsehood, a trend driven by populism in many countries. The response from the evidence community has in many cases been to produce even more evidence, which proves ineffective against populism. Despite this trend, evaluation systems are increasingly being strengthened within the political-administrative systems of many countries. This can be attributed in part to the influence of international actors and the crucial accountability function of evaluation.

This development of diverse NESs across multiple contexts has generated critical inquiries regarding optimal structural configurations, developmental trajectories and functional mechanisms, subsequently catalysing the development of systematic methodologies for describing and assessing NESs.

The systematic analysis and comparative examination of NESs represents a field of inquiry spanning approximately 25 years. The diagnostic assessment of NESs can serve various objectives: it can provide valuable empirical inputs for advancing scholarly research on NESs; it can establish baseline studies of NESs; and it can give direction on how to strengthen NESs.

As this discussion paper shows, different kinds of assessment offer important information on NESs, evaluation institutionalisation and evaluation culture. Many assessments of NESs offer a comparison of these systems via a set of indicators, providing valuable data from which conclusions can be drawn about the successful institutionalisation of evaluation.

Other assessments prioritise in-depth analytical approaches, thereby providing robust foundations for developing country-specific interventions to strengthen respective evaluation systems. The majority of existing assessments rely predominantly on subjective assessments from expert groups.

INCE represents a new approach that brings together two key elements: measuring evaluation system capacities using numerical data and building capacity through participation and networks.

This discussion paper presents a systematic comparison of five NES assessments along six criteria grouped in three categories: Category I - (1) the purpose and (2) development of the assessments; Category II - (3) the methodological design and (4) measurement process; and Category III - (5) the global coverage and (6) accessibility of instruments. The analysis focuses on comparative studies and instruments that have been applied in a minimum of ten countries and across at least three continental regions. INCE, being relatively new to the field and therefore not yet covered in academic literature, is a tool that shows promise for global evaluation system analysis. This paper seeks to introduce INCE to the scholarly community and evaluation capacity development (ECD) practitioners, thereby contributing to broader discourse on NES assessment methodologies.

Extremely important contributions have been made by various comparative assessments conducted at national and regional levels examining evaluation systems or policies (Chirau et al., 2020; Diwakar and Rao, 2023; Diwakar et al., 2022; Feinstein, 2014; Goldman et al., 2018; Kalugampititya, 2022; Toulemonde, 2000; Widmer et al., 2009) and by cross-country assessments (OECD, 2020). However, these studies fall outside the parameters of the present comparative analysis as they do not offer a global comparative perspective on evaluation systems and, therefore, are not suitable for a comparison with INCE.

Equally, the important work of Barbara Rosenstein, in cooperation with the Parliamentarians Forum for Evaluation and EvalPartners, on mapping national evaluation policies is a rich source of information. The mapping was implemented in 2013 (Rosenstein, 2013), 2015 (Rosenstein, 2015) and 2021 (Rosenstein and Kalugampitiya, 2021), and it provides comprehensive documentation of evaluation policy development trajectories. However, while evaluation policies constitute integral components of evaluation systems, Rosenstein and colleagues' thematically circumscribed focus and initial geographical concentration on Asian contexts preclude inclusion of this mapping in the present comparative analysis.

This discussion paper does not attempt to trace the historical evolution and development of scholarly work on NESs, but rather concentrates on the comparative examination of five diagnostic studies that are particularly relevant. For a detailed account of the history and development of the science of evaluation systems, we recommend Jacob's (2023) chapter titled "The institutionalisation of evaluation around the globe: Understanding the main drivers and effects over the past decades".

3.2 Brief introduction to the five assessments covered in the systematic comparison

The first systemic description and comparison of NESs around the globe was presented in the *International Atlas of Evaluation*, developed by Furubo et al. (2002), here referred to as Atlas 2002. This systematically describes and compares the evaluation systems of 21 countries and three international organisations. The work was implemented in 2001 and published in a book, with single chapters for each country and organisation, in 2002.

This work was updated 13 years later when Jacob, Speer and Furubo carried out a comparative analysis of 19 countries, all of them part of the first Atlas (Jacob et al., 2015). This 2015 work, Atlas Update, largely followed the methodology of Atlas 2002, also adding a longitudinal analysis. In contrast to Atlas 2002, Atlas Update was published as an academic article, focusing on description of developments between the 2002 assessment and the 2015 update.

Around five years later, the Department of Sociology and the Center for Evaluation (CEval) at Saarland University started the Evaluation Globe research project. To date, this project has analysed 38 NESs and six institutional evaluation systems in multilateral institutions across three regions of the world. The individual country studies were published in books for each region. The book on Europe, with 16 country studies, was published in 2020 (Stockmann et al., 2020); the book on the Americas, with 11 country studies, was published in 2022 (Stockmann et al., 2022a); the book on Asia-Pacific, with 11 country studies, was published in 2024 (Stockmann et al., 2024); and the book on Africa was not yet available when the analysis reported here was being carried out and, hence, does not form part of the comparison described.

In 2022, the Global Evaluation Initiative (GEI) developed the Monitoring and Evaluation Systems Analysis (MESA) diagnostic tool. The GEI is a global network of organisations and experts, operating under the umbrella of the World Bank, that aims to strengthen monitoring and evaluation systems. MESA is an important tool for the GEI's work. It allows for comprehensive description and analysis of a country's evaluation system, and serves as a basis for planning suitable ECD measures.

The most recent tool included in the comparison, INCE, was developed by different evaluation experts and ECD practitioners in Latin America and the Caribbean and is managed by the World Food Programme (WFP) and the German Institute for Development Evaluation (Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit, DEval). Its development dates back to 2017, and the first pilots were implemented three years later in five countries. INCE is currently (at the time of writing in 2025) used in 11 countries in Latin America, and it has also reached beyond Latin America, with six measurements being piloted in Africa and four more planned by the end of 2025 as well as one measurement successfully implemented in Asia (Mongolia). INCE data is presented in the form of a tree for each country. Until recently, no narrative reports had been produced on the individual evaluation systems measured; however, a pilot phase is currently underway to develop and standardise such reports, referred to as INCE-Briefs, as a complement to the diagnostic results. These briefs provide a narrative and contextualised interpretation of INCE findings, facilitating a more comprehensive understanding of each national evaluation system beyond the data itself and the tree visualisation.

While the five assessments all examine evaluation systems, they do so with somewhat different emphases: the authors of Atlas 2002 analyse evaluation culture, the authors of Atlas Update describe the analysis of both evaluation culture and institutionalisation, the Evaluation Globe authors analyse evaluation systems according to a specifically defined concept, and the MESA and INCE authors focus on the capacities of evaluation systems. Despite these differences in focus, the following discussion treats these assessments collectively, emphasizing their shared objective of examining evaluation systems rather than maintaining strict distinctions between their respective focal points

4. PURPOSE AND DEVELOPMENT OF THE ASSESSMENTS

Improving our understanding of NESs requires an effective link between scientific analysis and insights and their practical application, which should then be assessed and fed back into the research process. The five assessments examined in this discussion paper serve dual functions – providing academic contributions while also aiding comprehension and offering practical support for the enhancement of NESs. The main purposes underlying each assessment fundamentally determine its development, implementation approach and utilisation patterns. This has resulted in substantial variation across the five analytical works examined

4.1 Purpose of the assessments

Atlas 2002 seeks to describe the current situation of evaluation systems around the globe, describing relevant developments. It is available in book form, with one chapter devoted to each evaluation system. A second purpose of the book is to explain these developments, considering "what forces are affecting the contour of evaluation in different national contexts and what consequences these forces will have on the diffusion of evaluation" (Furubo et al., 2002: 1).

Atlas Update is a scientific article consisting of 31 pages, so the information it contains is much more condensed. It seeks to build on Atlas 2002 and to recognise the diversity of NESs. The authors describe its purpose as follows: "The development of evaluation culture does not follow a one-dimensional model. This makes developments empirically difficult to capture and these challenges are further compounded by the varied historical roots for governance. Being conscious of these complexities and challenges, we sought to analyse the diverse forms and functions of evaluation culture in 19 OECD countries around the world" (Jacob et al., 2015: 20). With its longitudinal analysis, Atlas Update also aims to analyse what happened since the first Atlas analysis, thus identifying and describing major trends in the institutionalisation of evaluation as well as components of strong NESs.

Evaluation Globe has, at the time of writing, published three books with a fourth due – one for each region covered. Accordingly, compared to the other works described here, the information provided is significantly more comprehensive. Evaluation Globe pursues a similar purpose to Atlas 2002 and Atlas Update. According to Stockmann et al. (2022b) the Evaluation Globe seeks to describe the institutionalisation of evaluation across different countries worldwide and subsequently conduct a comparative analysis of its historical development, current state, and future perspectives. It is aimed explicitly at an interdisciplinary audience (researchers, politicians and administrative staff) with the goals of improving understanding of the institutionalisation of evaluation across the world and guiding actions to strengthen evaluation systems.

MESA studies are not intended as ends in themselves, but lay the groundwork for planning ECD activities. The purpose is described on the GEI home page as follows:

> *The MESA is a diagnostic tool that guides country stakeholders (e.g., government entities, evaluation professionals, civil society) in gathering, structuring and analysing information on the current capacity of their country's M&E [monitoring and evaluation] ecosystem. It helps identify what is working well and what needs to be improved, informing capacity-development strategies meant to strengthen the economic, political, and social context that enables M&E to flourish.*

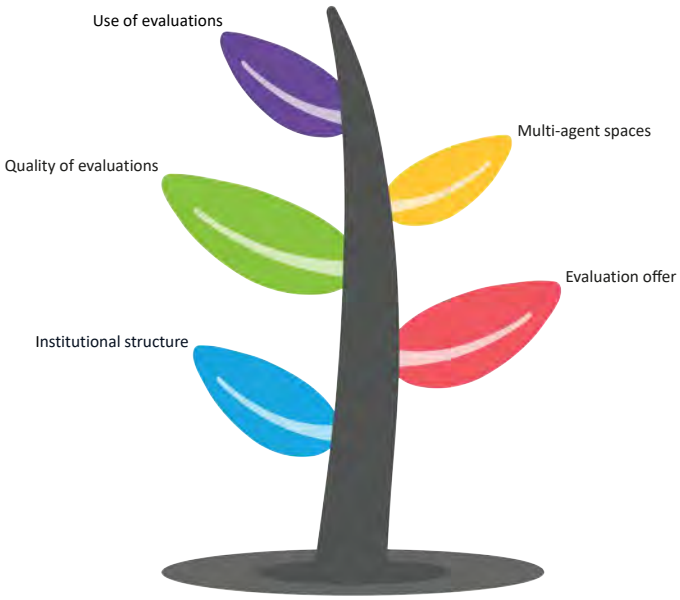
A MESA is not an end in itself, but rather a means to gather, structure, and analyze information to inform and shape improvements to a country's M& systems (GEI, n.d.)

MESA is directed at a diverse audience and covers not only evaluation but also monitoring. It is the only assessment covered in this discussion paper that also analyses national monitoring systems. Due to the flexible implementation of MESA, the studies can also focus on those parts that are of special interest to a country's NES stakeholders (mainly the national evaluation unit), making it even more useful for these systems. MESA represents a key tool for GEI's work to strengthen NESs, and thus MESA final reports are designed to provide the basis for national ECD strategies.

The primary objective of INCE is to enhance NESs through sustained knowledge exchange, peer-to-peer learning mechanisms and network development. INCE provides a structured and comparable measurement of national evaluation capacities across countries. In doing so, it references an ideal NES as its conceptual framework. This ideal system comprises five distinct dimensions (see Figure 1), each representing a critical component of a well-functioning evaluation system. By measuring these dimensions systematically, the index enables a comprehensive understanding of how closely a country's actual evaluation system approaches the theoretical ideal, thereby identifying both strengths and areas for potential development within the NES.

A tree metaphor (shown in Figure 2), with distinct branches representing the five NES dimensions measured through INCE, serves to illustrate that the index prioritises diversity over inter-country comparison. This visualisation strategy emphasises variation and complementarity rather than competitive positioning. It demonstrates which countries possess particular strengths across specific dimensions, thereby facilitating identification of potential knowledge providers and recipients within the peer-to-peer learning network.

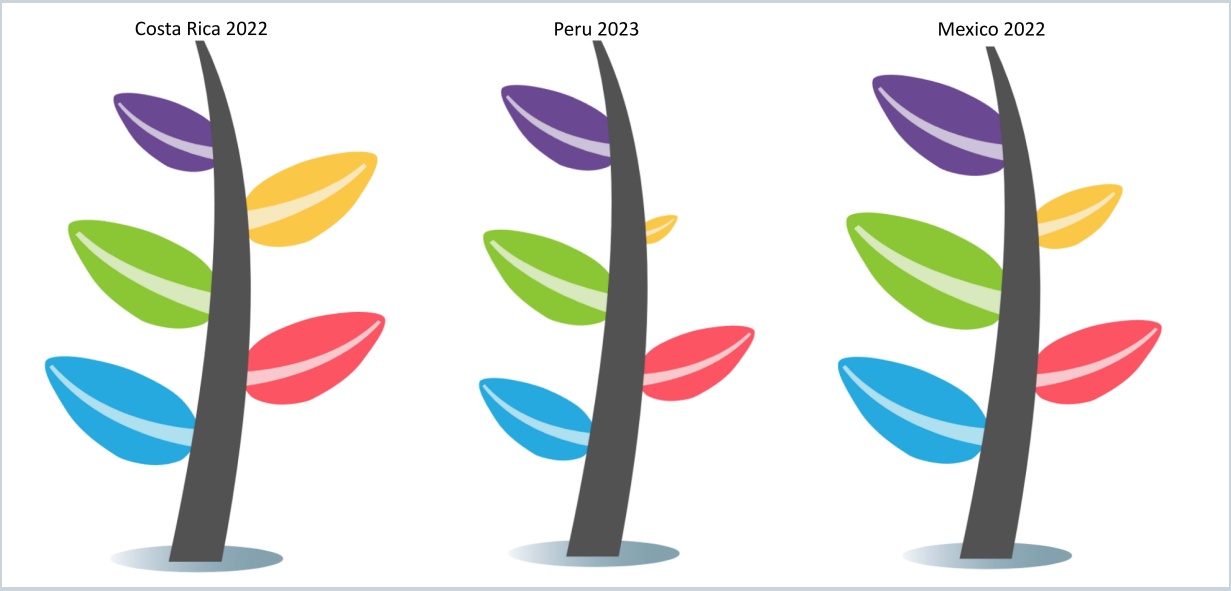
Figure 2 Tree representation of INCE dimensions



Source: DEval, own visualisation

This design feature reflects that the collaborative learning paradigm was systematically integrated into INCE's methodology from its conceptual inception, ensuring that the measurement framework directly supports the assessment's capacity-building objectives rather than merely documenting system characteristics. Furthermore, collaborative networks have been a constant feature of INCE, from the instrument's development to the data collection and interpretation and use of results, demonstrating its strong emphasis on peer learning. In addition, the regular application of the instrument – every two years – allows for monitoring the evolution of NES capacities over time.

Figure 3 Tree representation of INCE dimensions



Source: DEval, own visualisation

Costa Rica achieved high scores on the multi-agent dimension (see the Costa Rica tree for 2022), which is an area that Peruvian NES officials sought to strengthen within their framework (see the Peru tree for 2023).

To facilitate this knowledge transfer, representatives from Costa Rica's Ministry of National Planning and Economic Policy (Ministerio de Planificación Nacional y Política Económica, MIDEPLAN) were invited to Peru's workshop on utilizing INCE results, where they shared their experiences of and best practices for multi-agent engagement.

Similarly, Costa Rica invited representatives from Mexico's National Council for the Evaluation of Social Development Policy (Consejo Nacional de Evaluación de la Política de Desarrollo Social, CONEVAL) to their 2023 workshop, as they wanted to learn how Mexico performed so well on the utilisation dimension (see the Mexico tree for 2022).

4.2 Development of the assessments

Atlas 2002 and the subsequent Atlas Update were developed by evaluation scholars from the Global North with extensive academic credentials in evaluation theory and substantial practical evaluation experience. The individual country chapters in the Atlas 2002 book were developed collaboratively with national and regional evaluation experts, ensuring contextual validity and integration of local expertise. The other three assessments presented in this discussion paper utilised Atlas 2002 and Atlas Update as foundational references during their development phases.

Evaluation Globe was developed and edited by academics and evaluation experts from the Global North, but the four volumes (including the one on Africa) were compiled in close cooperation with more than 100 evaluation experts from all over the world, who are named authors of the individual chapters. The compilation of Evaluation Globe can therefore be seen as a global exercise.

The MESA structure was developed by evaluation experts and ECD practitioners. MESA studies are always carried out by national and international evaluation experts, and the MESA Guidance Note declares country ownership as one of the five principles underpinning the MESA process (GEI, 2022: 17). Indeed, MESA studies are only carried out in those countries which state a clear demand for a study, and all MESA studies are co-led by a national institution, such as a line ministry.

Like MESA, INCE was developed in the context of development cooperation. In contrast to the other assessments presented, INCE is a joint initiative of a large and diverse group of stakeholders. The idea to create an index to measure national evaluation capacities was born in a gathering of evaluation stakeholders in Mexico in December 2017. Shortly after, a working group was formed and INCE was developed in a participatory process that included over 20 actors from Latin American countries. These actors represented national evaluation units and academia as well as multilateral, regional and bilateral institutions. Two organisations manage the INCE process: the WFP and DEval. A list of all actors involved in the development of INCE can be found in Annex 1.

INCE can only be implemented where a national evaluation unit exists and is willing to assume leadership on the application of the instrument and the subsequent use of results. Accordingly, INCE is implemented by national evaluation units in Latin America, Africa and Asia. However, other actors were central to the spread of INCE, establishing the necessary relationships and supporting the first measurements. Over the last years, the Inter-American Development Bank (IDB) and the Center for Learning on Evaluation and Results for Latin America and the Caribbean (Centro para el Aprendizaje en Evaluación y Resultados de América Latina y el Caribe, CLEAR-LAC – a GEI implementing partner) engaged very actively in INCE measurements across Latin America, contributing significantly to its development. In Africa, the WFP and DEval worked under the leadership of the African Evaluation Association (AfrEA) with the support of the individual national evaluation associations in the six African pilot countries (Benin, Republic of the Congo, hereafter Congo-Brazzaville, Ghana, Morocco, South Africa and Tanzania). In Asia, INCE was introduced in Mongolia, in close cooperation with the national evaluation association, and Sri Lanka, in cooperation with UNICEF.

Some of the evaluation experts who wrote country chapters for the Evaluation Globe compendium on the Americas were involved in the development of INCE. Equally, the developers of MESA had regular exchanges with the INCE team. Both teams – INCE and MESA – are keen to make best use of the combination of the two diagnostic tools.

4.3 Summarizing analysis of the purpose and development of the five assessments

The analysis reveals a fundamental distinction between **assessments designed primarily for (scholarly) discussion and those developed for direct practical application** (see Figure 4). Atlas 2002, Atlas Update and Evaluation Globe offer valuable information for the ECD community. These works function as stand-alone academic resources that contribute to scholarly discussion on evaluation systems, but they do not have specific objectives indicating how they are to be used. The MESA studies are directed towards national evaluation units, and they consistently generate concrete outputs in the form of national ECD plans, developed through collaborative processes between national stakeholders and GEI technical staff. This approach facilitates direct translation of assessment findings into actionable strategic plans.

Figure 4 Purpose of the assessments[3]

Source: DEval, own visualisation

INCE frequently produces important input for ECD measures, yet its primary contribution lies in facilitating sustained dialogue between national and international actors regarding NES characteristics and enhancement strategies. The assessment process generates ongoing collaborative engagement that extends beyond initial measurement activities. A recent pilot measurement conducted in Congo-Brazzaville exemplifies these dynamics, which resulted in the establishment of formalised regular exchange between parliamentary representatives and the national voluntary organisation for professional evaluation (VOPE). These collaborative outcomes are intrinsically linked to INCE's distinctive measurement processes, which are examined in detail in the subsequent section of this discussion paper.

The way assessments were developed and the participation of diverse stakeholders in the assessments vary greatly. The Atlas publications, developed by scholars from the Global North in consultation with national experts, maintain academic rigor while ensuring contextual validity. Evaluation Globe expanded this approach by involving over 100 global experts as chapter authors, creating a more comprehensive global exercise. MESA and INCE represent a paradigm shift towards participatory development: on the GEI home page, country ownership is emphasised as a core principle of MESA (as noted in Section 4.2); and from its inception, INCE's three-year participatory development process has involved over 20 actors across multiple institutional categories.

Another difference between the assessments is linked to measurement frequency and data utilisation, which manifests clearly in measurement approaches. Atlas, Evaluation Globe and MESA provide in-depth, irregular assessments that generate comprehensive country reports suitable for academic analysis and strategic planning. INCE operates differently, conducting systematic biennial measurements that produce extensive datasets presented through condensed graphical representations, prioritizing ongoing monitoring over detailed narrative analysis. This reflects INCE's primary purpose of facilitating continuous dialogue as opposed to producing stand-alone analytical products.

[3] The analysis carried out in this discussion paper forms the basis for classification of the individual assessments. The table in Annex 2 provides more detailed information on the assessment of the categories of direct contribution to strengthening NESs and value for scholarly discussion.

The most significant differentiation emerges in post-assessment outcomes. Atlas (2002 and Update) and Evaluation Globe function as information resources and do not have embedded mechanisms for sustained engagement. The MESA studies are used for the formulation of ECD plans. INCE's participatory development approach, however, has generated a dynamic network of evaluation professionals who utilise the framework as both an information exchange platform and a strategic planning tool for developing concrete interventions to strengthen NESs. This network effect demonstrates the integration of assessment methodology with capacity-building objectives, creating a system where measurement activities contribute to system enhancement through collaborative knowledge sharing and coordinated action planning.

While INCE's participatory methodology generates substantial ownership and acceptance, it requires significantly more coordination resources and longer development cycles compared to expert-driven approaches like that of Atlas. The time frame from the start of development to the first pilot measurement was three years. In addition, even minor modifications and adjustments often require extensive coordination processes with correspondingly longer implementation times. This comes at a cost: while INCE is very well known in Latin America, where it was developed and first implemented, the actors involved have not yet invested enough resources to make INCE known and usable more broadly. In its (further) development and implementation, INCE operates in an area of tension between global participation to ensure quality and acceptance on the one hand and the associated time and resource-related challenges on the other.

5. METHODOLOGICAL DESIGN AND MEASUREMENT PROCESS

The institutionalisation of evaluation always adapts to the context of a country, which is why evaluation systems will always differ and will continuously develop (Gaarder and Briceno, 2010; Pérez Yarahuán and Trujillo, 2015). At the same time, globally valid criteria are required to be able to assess and compare the systems – not necessarily in a competitive sense, but rather as a basis for mutual learning and cross-country exchange – though these criteria are not final and irrevocable. It can be assumed that the discourse on NESs will become more global in the coming years and that this will also bring new perspectives and more developed understanding of them, including the criteria that form the basis of assessments and the indicators that best describe NESs.

In their work on the relationship between evaluation culture and good governance, Dahler-Larsen and Boodhoo (2019) describe two central points for the description, measurement and comparison of evaluation systems. First, it cannot be a question of these systems becoming ever stronger/more mature; at some point, the system cannot become stronger through further institutionalisation but should adapt to the changed conditions of a society through qualitative adjustments (for example, through different evaluation approaches and new methods). The second important point is that indicators for measuring a NES are always based on the idea of an ideal system. However, these ideas differ from region to region and sometimes from state to state or even from stakeholder to stakeholder, a dilemma difficult to deal with in systematic comparisons between evaluation systems in different parts of the world. The compared works in this discussion paper deal with this differently, as described in the subsequent subsections 5.1 and 5.2

5.1 Methodological design and structure

In Atlas 2002, the analysis and assessment of the maturity of evaluation cultures is based on nine indicators, covering the demand and supply sides of evaluation: (1) evaluation takes place in many policy domains; (2) there is a supply of evaluators specializing in different disciplines; (3) discussions and debates fuel a national discourse regarding evaluation; (4) a national evaluation society exists; (5) institutional arrangements exist in the government for conducting evaluations and disseminating their results; (6) institutional arrangements exist in parliament for conducting and disseminating evaluations; (7) pluralism exists within each policy domain; (8) evaluation activities occur within the supreme audit institution; and (9) evaluations focus not just on inputs/outputs but also on outcomes.

The indicators were developed based on two main factors showing the maturity of an evaluation culture: the degree to which evaluation is institutionalised and the "actual spread and pluralism of the evaluative culture and its openness to new ideas and impulses" (Furubo et al., 2002: 7). The scores for each country are the product of an interactive process between the authors of the individual country chapters and the editors of Atlas 2002.

Atlas Update uses the same indicators as Atlas 2002 and the same scale, but it puts its assessment on a broader basis by conducting a survey with evaluation experts from the countries analysed. Four to five experts from three different groups – public sector, private sector and academia – worked on each country assessment, and a total of 78 people took part in the survey.

It can be assumed that the experts whose judgments were used for the scores in Atlas 2002 and Atlas Update used the data sources available to them (such as literature, country statistics and conversation with colleagues) for their respective assessments.

Evaluation Globe is based on case studies for each country and institution. Each case study followed a predefined analytical framework, covering three dimensions of evaluation systems: the political system (institutional structures and processes), the social system (societal dissemination and acceptance of evaluation in society) and the system of professionalisation (professionalisation of evaluations). Guiding questions were derived from the analytical framework. Each case study was elaborated by an evaluation expert from the respective country, who was asked to not only describe their perception of the system but also back it up with as much data as possible using literature and document analysis as well as around five interviews with other evaluation experts. The guiding structure (framework) of the case studies along the three dimensions allows for rather detailed comparisons between the different countries, institutions and regions. The individual country chapters of the Evaluation Globe volumes describe the national level primarily, but also highlight important developments at subnational and sectorial levels.

MESA is structured around five parts: (1) the background of the country and its current status in relation to M&E; (2) an overview of the country's overall public sector management capacity; (3) the monitoring and reporting systems; (4) the evaluation systems; and (5) the findings, conclusions and recommendations. MESA does not measure fixed indicators, but recommends using recognised indicators for preparing the country profiles (socio-political background), such as the United Nations Development Programme's (UNDP's) human development indicators. MESA studies provide a comprehensive analysis of a country's evaluation system, embedding it within the broader national context – an approach that is highly valuable for identifying strengths and weaknesses. As a diagnostic tool for M&E systems, MESA goes beyond evaluation by also incorporating key aspects of monitoring into its analysis. Also, MESA is a flexible tool and can be applied to different kinds of M&E systems – those of institutions, sectors, cities or countries. MESA's aim is not to compare the different systems, so all five components do not need to be addressed. Therefore, it can be adapted to the demands and needs of those working on strengthening the relevant system, and this flexibility makes it very useful for ECD practitioners.

INCE describes NESs and measures their capacities along five dimensions: (1) institutional structure, (2) evaluation offer, (3) quality of evaluations, (4) multi-agent spaces and (5) use of evaluations (see Figure 2). These five dimensions are operationalised in 18 subdimensions, 38 variables and 74 indicators. INCE is a summative non-weighted index with values ranging from 0 to 10. Data is collected through an online survey completed by institutions from three groups: the national evaluation unit as governing entity of the NES, other actors in the system and external actors. The integration of divergent perspectives into the measurement methodology is based on systems theory insights, which postulate the existence of multiple, partially contradictory viewpoints within complex systems. According to systems theory, only the totality of these heterogeneous perspectives constitutes representation of a complete system, as individual perspectives can only capture partial system segments (Bohr, 1928; Keating et al., 2020; Klier et al., 2022). To operationalise this theoretical premise and to facilitate learning between different INCE measurements, a systematic categorisation of the identified perspectives was undertaken. The viewpoints that are inherently present in every NES were organised into a three-tier classification matrix (see Table 1), which enables standardised data collection and facilitates cross-case reflection and perspective-specific analysis.

Table 1 The three groups of NES stakeholders answering the INCE survey

Governing entities (national evaluation units)	Other NES actors	External actors
The governing entity, typically within a government ministry or planning agency, coordinates and oversees the NES and sets evaluation standards.	Other NES actors consist of government agencies, non-governmental organisations and academic institutions engaged in the implementation and utilisation of evaluations.	External actors include international donors, academic institutions, consultancy firms and multilateral organisations that support evaluations independently of the national system.

Source: DEval, own visualisation

5.2 Measurement process

Each of the assessments work with experts from the respective countries. For Atlas 2002, Atlas Update and Evaluation Globe, data was collected and analysed in close cooperation with national evaluation experts. The editors of the respective assessments shared the data collection and analysis framework, and national experts gathered the data and prepared the country chapters. Final data analysis was done jointly. This works in a similar way for MESA. The authors of a MESA study collect and analyse data corresponding to the MESA framework found on the GEI home page. The MESA studies are co-led by national institutions, which are involved very actively in the process of MESA development. In order to collect the relevant data for studies, a series of meetings takes place, connecting the different actors involved in the NES and offering space for them to discuss the strong and weak points of the system. These spaces of interaction and exchange are important for strengthening evaluation systems, so the process of MESA development is one of its most valuable aspects.

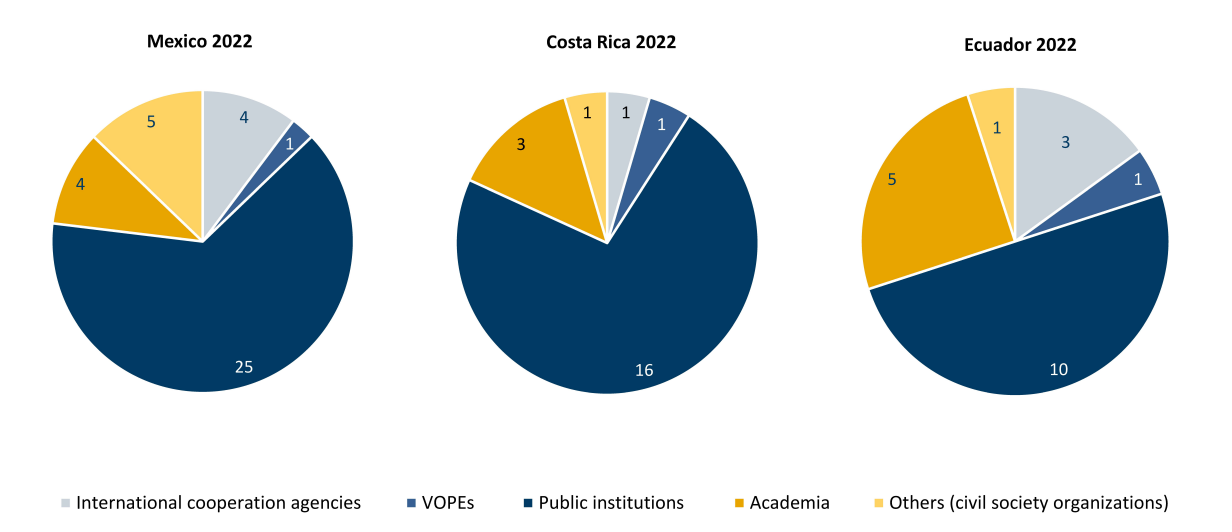
Among the instruments considered, INCE – being an index – stands out for offering a more detailed measurement process. In contrast, the other diagnostic tools are primarily expert- or consultant-led studies and have varying degrees of local stakeholder involvement, taking place through interviews or consultation processes. As such, only INCE allows for a meaningful analysis of its measurement process, which is why this section focuses on this instrument.

INCE assessments are characterised by a relatively short implementation time frame, with standard measurement cycles typically completed within a three-week period. Data is collected through an online survey, the administration of which comes under the institutional oversight of the designated national evaluation units (referred to as governing entities), which receive technical support from INCE specialists throughout the assessment process.

Based on their knowledge and central role in the NESs, the national evaluation units compile a list of key informants, organised according to the three categories in Table 1, to complete the online survey. The INCE team can provide guidance based on lists developed in other countries.

The INCE measurement process incorporates all relevant stakeholders within NESs, thereby generating substantial ownership and commitment among national actors. The distribution of stakeholder groups is shown in Figure 5 for the three countries that had their first measurement in 2022, after having participated in the INCE pilot in 2020.

Figure 5 Stakeholder distribution in INCE 2022 measurements



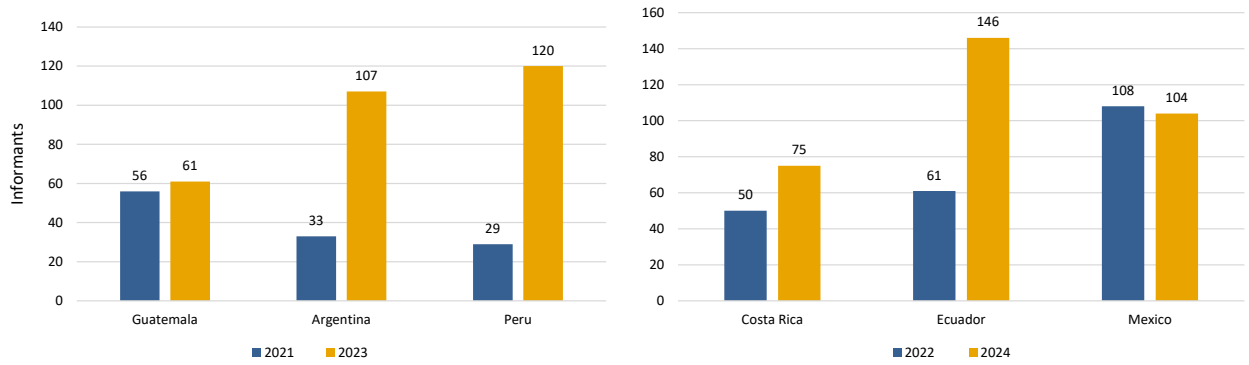
Source: DEval, own visualisation

The 2022 measurement cycle in Mexico, Costa Rica and Ecuador illustrates the engagement of a broad range of national actors in the development of evaluation systems. In each country, various institutional actors – including public institutions, academia, civil society organisations and international cooperation agencies – contributed to the process. With 39 participating institutions in Mexico, 22 in Costa Rica and 20 in Ecuador, the data reflects both a high level of national participation and the institutional diversity that supports inclusive implementation and strengthens national ownership of evaluation efforts.

This methodology results in considerable inter-country variation in sample composition and sise, and participation rates also demonstrate substantial heterogeneity across implementation contexts. Empirical data illustrates this variability: the 2025 Tanzania assessment engaged 774 respondents representing 122 institutional entities, whereas the 2021 Guatemala implementation comprised 56 participants from 14 institutional units. This reflects the contextual adaptation of sampling frameworks to national evaluation ecosystem characteristics.

This measurement process embodies fundamental design principles of INCE, specifically national ownership and participatory network integration. In general, one can see a strengthening of network integration, based on the number of people participating in the online survey. As Figure 6 depicts, this number in most cases increases from the first to the second measurement, resulting in more people discussing the NES. In most cases it is these same people and entities that are invited to a workshop around the INCE results (see below), which further strengthens the evaluation network in the country.

Figure 6 Change in the number of informants for INCE measurements



Source: DEval, own visualisation

However, this methodological approach introduces inherent limitations regarding cross-national comparability and temporal consistency, as measurements are conducted by varying personnel across implementation cycles – a recognised challenge endemic to longitudinal research design (Hollstein, 2021).

To mitigate these methodological constraints, the INCE working group strategically designed the data collection framework with an institutional rather than an individual focus. The protocol explicitly encourages collective institutional reflection and deliberative response formulation, leveraging the distributed expertise of evaluation personnel within participating organisations. This collaborative approach serves to attenuate individual subjectivity while enhancing data validity through triangulation of institutional knowledge.

While this methodology may constrain strict cross-national and temporal comparability, the approach prioritises internal validity and organisational learning outcomes over standardised measurement conditions. This is a recognised trade-off in composite index construction (Nardo et al., 2008; Saltelli, 2007). The deliberate emphasis on institutional-level analysis strengthens the reliability of collected data through reduced measurement error associated with individual response bias.

Comprehensive analytical guidelines governing INCE implementation protocols are publicly accessible through the official INCE platform, ensuring methodological transparency and replicability across assessment contexts.

While numerical results constitute a fundamental component of the index, INCE distinguishes itself from the other four assessment instruments examined in this discussion paper through its extension beyond conventional measurement, analysis and dissemination protocols. A critical dimension of the INCE process commences subsequent to these standard procedural phases.

Following each measurement process, national evaluation units retain discretionary authority to implement INCE utilisation workshops. Almost all evaluation units choose to do so. These structured workshops facilitate systematic national stakeholder engagement, wherein national evaluation actors conduct comprehensive results analysis and strategic planning exercises. The workshops frequently incorporate cross-national peer-to-peer learning through the participation of evaluation units from other countries, specifically targeting capacity enhancement in dimensions identified as weak.

The workshop results are regularly used for the formulation of national ECD plans owned by the national evaluation units. This process reflects very well Mackay's reflection on the importance of cooperation between the actors involved in evaluation systems: "Achieving real coordination among all these actors is typically not easy, so any process such as preparation of an M&E diagnosis provides one opportunity to get the key stakeholders to talk to each other about M&E and to attempt to reach some agreement on what to do about improving the government's approach" (2007: 68).

National ownership of INCE is demonstrated through the adoption and adaptation of the instrument by various countries for processes related to their domestic evaluation systems. This provides evidence of INCE's practical applicability as an assessment tool and indicates its successful institutionalisation within NES frameworks. This phenomenon suggests that INCE has transcended its role as an external assessment mechanism to become an integral component of national evaluation infrastructure. For example:

- In **Costa Rica**, INCE is used to monitor national evaluation policy.
- In **Colombia**, INCE is used as a baseline and to monitor the country's national evaluation strategy (Departamento Nacional de Planeación, Consejo Nacional de Política Económica y Social, 2022).
- In **Peru**, INCE is integrated into the National Strategic Development Plan (Plan Estratégico de Desarrollo Nacional 2050, published in 2022) as an indicator measuring the NES.
- **Dominican Republic and Ecuador** use INCE as an instrument to harmonise ECD engagement.

In Latin America, INCE is also the basis for harmonisation in the ECD space. Several ECD actors meet monthly to inform each other about ongoing and planned ECD activities and to join forces to implement selected ECD activities.[4]

As noted earlier, at the time of writing in 2025, INCE operates in the described way exclusively within the Latin America and the Caribbean region, with 11 countries in this geographic area having adopted the instrument. In 2025, six pilot implementations were initiated across African countries (Benin, Congo-Brazzaville, Ghana, Morrocco, South Africa and Tanzania), with participating experts maintaining weekly coordination meetings to ensure methodological consistency and knowledge transfer.

This African expansion has catalysed the emergence of a structured network encompassing national evaluation units and VOPEs from the six participating countries. This network benefits from pre-existing evaluation networks, especially AfrEA and its efforts to institutionalise evaluation in countries across the region.

The African continent presents significantly greater geographic scale and socio-cultural diversity compared to the Latin American context. While Latin American INCE processes predominantly use Spanish, the African working group necessitates multilingual coordination in English and French. While these are official languages in the respective countries, it should be noted that they may not constitute the primary languages for all stakeholders engaged in national INCE processes.

The linguistic complexity anticipated in African implementation foreshadows similar challenges for potential Asian regional expansion. Beyond linguistic considerations, additional contextual adaptations will likely be required, and INCE's transferability and methodological consistency across diverse regional contexts remains to be proven.

Current initiatives are under way to systematically adapt INCE for implementation within African and Asian regional contexts. This adaptive approach is methodologically grounded in the recognition that index construction is inherently predicated upon underlying theoretical models – in this case, conceptualisations of optimal evaluation system architecture – which may exhibit regional variation based on contextual factors, institutional frameworks and cultural paradigms.

The regional adaptation process acknowledges that standardised assessment criteria developed within one geographic context may require calibration to ensure validity and relevance across diverse evaluation systems. This methodological flexibility ensures that INCE maintains its analytical rigor while accommodating the heterogeneous characteristics of evaluation systems in different regional contexts.

[4] DEval, the WFP, CLEAR-LAC, the Office of Evaluation and Oversight of the Inter-American Development Bank (OVE-BID), UNICEF, UN WOMEN, the GEI and the United Nations Economic Commission for Latin America and the Caribbean (Comisión Económica para América Latina y el Caribe, CEPAL) participate regularly in these meetings.

In the adaptation process so far, it has become clear that in Latin America great importance is attached to the network character of an evaluation system (a dimension that is not present in the other assessments presented in this discussion paper, but which responds to the work around evaluation and systems theory (Hummelbrunner, 2011; Klier et al., 2022; Williams and Imam, 2006), and in Africa, an evaluation system might be considered particularly good if local evaluation approaches are anchored in it. The process of adapting INCE to the African context has only just begun, and it remains to be seen whether this will lead to the inclusion of an additional dimension.

An attempt is being made to establish a standardised core index that maintains consistency across all regional implementations while also incorporating region-specific components where methodologically justified and contextually relevant. As stated above, compared to the diverse contexts found within Asian and African regions, the countries within Latin America and the Caribbean demonstrate considerably greater homogeneity across linguistic, social and cultural dimensions, including in terms of religion. This fundamental difference in regional coherence suggests that singular, unified models for Africa and Asia may not be methodologically appropriate or practically feasible.

Additionally, thematic modules are under consideration (including gender and climate evaluation dimensions), though such expansions would necessitate extended measurement periods and may reflect the priorities of international development stakeholders rather than national evaluation actors.

However, based on the experiences from INCE pilot measurements completed to date in Asia and Africa (Mongolia, Congo-Brazzaville, Ghana, South Africa and Tanzania), substantial modifications to the underlying theoretical model appear unnecessary. Instead, primary adaptation requirements centre on implementation processes, particularly data collection methodologies and the provision of comprehensive preparatory training for local implementation teams. Also, given INCE's reliance on digital platforms, ensuring technological accessibility and connectivity represents a critical consideration for successful implementation across many African contexts.

Despite the methodological challenges the adaptation process might bring, it must also be said that the process of reflection on possible adaptations of INCE and what constitutes a perfect evaluation system is very valuable for the development of national evaluation capacities, and this also responds Dahler-Larsen and Boodhoo's (2019: 281) note on the challenge to measure evaluation systems with indicators that are fit for each context, as described at the beginning of the section.

The adaptation process thus serves a dual function: ensuring contextual relevance of the assessment instrument while simultaneously fostering critical reflection on evaluation systems among national stakeholders. When carried out in a participatory manner, as seen in several Latin American countries, the adaptation process also contributes to building or strengthening regional collaboration networks and reinforces national ownership of the instrument. This approach transforms potential methodological challenges into opportunities for capacity building and theoretical refinement of evaluation system conceptualisation.

Apart from challenges around the application of INCE to other regions, the current INCE methodology exhibits an analytical constraint for comprehensive evaluation system assessment: INCE measurements lack contextualisation within specific national governance structures, such as political system configurations. This represents a significant methodological gap, as the development trajectory and complexity of NESs vary substantially between federal and unitary state structures. Federal systems typically present greater institutional complexity and coordination challenges for evaluation system development compared to unitary governance arrangements.

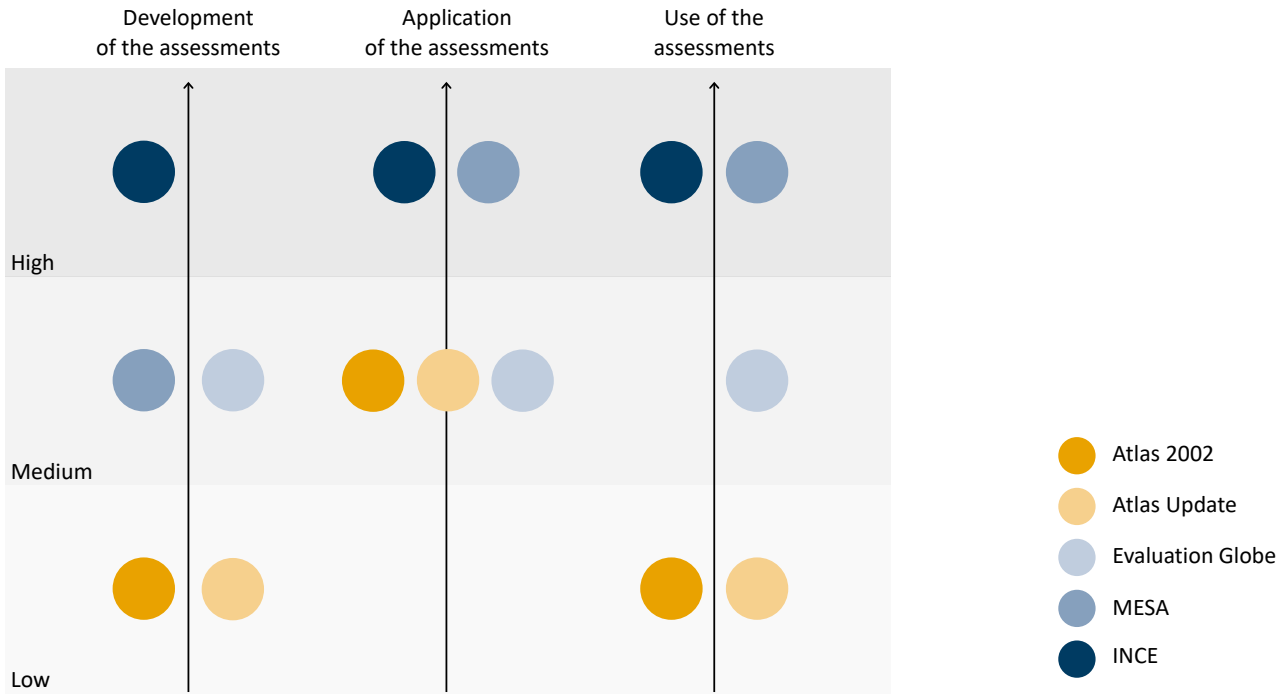
Currently, INCE does not systematically account for these fundamental structural differences, potentially limiting the validity of cross-national comparisons.

5.3 Summarizing analysis of the methodological design and measurement process

The examined assessments demonstrate comprehensive representation of NESs, exhibiting complementary characteristics while maintaining distinct analytical emphases without methodological contradiction. Each assessment contributes unique perspectives to the assessment of evaluation system architecture, resulting in a coherent body of diagnostic tools.

Atlas 2002, Atlas Update and Evaluation Globe offer valuable analyses of the development of evaluation systems. These kinds of analysis are a necessary input for further development of the field around strengthening evaluation systems. MESA and INCE do not offer this comparative analysis and the resulting explanations for why groups of countries have strong or weak evaluation systems. The data contained in both MESA and INCE would allow for this, but apparently this has not been used in such a way so far, which might be due to insufficient sharing of data with interested parties. INCE distinguishes itself through the substantial leadership that local stakeholders assume throughout the development, application, and utilisation of the instrument, see Figure 7.

Figure 7 Level of leadership of the assessments among national stakeholders of the countries assessed[5]



Source: DEval, own visualisation

INCE, Atlas 2002 and Atlas Update fall relatively short when it comes to the subnational level, where a major part of policies are developed and implemented in many countries. This is a major shortcoming. An INCE working group is currently developing an adaptation of INCE to take the subnational level into account, and although a beta version of INCE Subnational is now ready and will be tested in 2025 in both Peru and the region of Navarra (Spain), it can be assumed that this will take quite some time to complete. Evaluation Globe and MESA integrate this important perspective, and MESA even puts a focus on this perspective when demanded (for example, in the MESA elaborated for Paraguay).

[5] The analysis carried out in this discussion paper forms the basis for the classification of the individual assessments. The table in Annex 4 provides more detailed information on the level of leadership undertaken by national stakeholders.

These gaps, the analysis of evaluation use and the consideration of subnational levels represent important areas for future methodological development, as comprehensive evaluation system assessment requires systematic attention to both use patterns and multilevel governance structures that characterise contemporary evaluation systems.

An additional area for future methodological development concerns the integration of artificial intelligence (AI) technologies within NES assessments. Current assessments rely exclusively on conventional data collection and analytical methodologies, with AI applications remaining absent from existing diagnostic tools. This could change in the future, as Eckhard et al. (2024) are showing in their work on the analysis of institutional evaluation systems with the use of AI.

The integration of AI technologies could potentially enhance both the efficiency and the analytical depth of evaluation system assessments through automated data processing, pattern recognition and predictive modelling capabilities. Such technological integration represents a significant opportunity for methodological advancement in the field of evaluation system assessment, warranting systematic exploration and empirical validation in future research endeavours. In fact, some countries applying INCE have requested a narrative report to accompany the diagnostic results. A pilot phase is currently underway in Mongolia, Chile, Guatemala, and South Africa to develop these reports, with the aim of standardizing this practice in the coming months. The use of AI could streamline this process without significantly increasing the resources and time required for each assessment.

6. GLOBAL COVERAGE AND ACCESSIBILITY OF INSTRUMENTS

A growing understanding of NESs requires in-depth analyses of individual evaluation systems as well as overviews and comparisons. Comparative work is particularly easy to understand and methodologically feasible if the countries are largely similar. However, comprehensive global understanding of NESs also requires data on their differences and, therefore, data from as many countries as possible. Ideally, this data would be accessible for as many interested people as possible, scholars and practitioners alike. Only when it is used can the available data turn into a valuable resource.

Naturally, evaluation systems from different regions of the world differ greatly. This is due to their specific development and history as well as external factors. Just as institutional evaluation systems are shaped by their organisational culture Raimondo (2018), NESs are shaped by the respective country's political system, culture and context (Bustelo and Jacob, 2025; Dahler-Larsen and Boodhoo, 2019; Lahey, Robert, 2025). While evaluation systems in advanced economies started to develop as early as the 1960s, in Latin America this only happened 10 years later, while the first evaluation systems in African countries emerged roughly 20 years ago (the first one being Benin in 2007, followed by Uganda and South Africa in 2011; (Goldman et al., 2018)). This development is relevant because in most cases, the maturity of an evaluation system is reflected in its functionality. In many European countries, the institutionalisation of evaluation has been driven by the EU (Martinaitis et al., 2019; Pattyn et al., 2017), while evaluation practice in many countries in Africa, Asia and Latin America is often still dominated by bilateral and multilateral organisations. It is somewhat paradoxical that NESs in OECD countries are less developed than those in developing countries. While several countries in the Global South are at least attempting to set up an evaluation system covering all ministries, in many OECD countries, NESs are limited to single sectoral ministries: development cooperation is by far the most advanced sector, followed by education, health and labour.

Both internal factors – like culture, economic strength and political systems – and external factors – like the influence of multilateral organisations – are decisive for the development of NESs and one reason for their great heterogeneity. Ideally, assessments of NESs reflect this heterogeneity by examining many countries, representing a degree of diversity.

6.1 Accessibility

In order for assessments to be used to strengthen NESs around the world, their accessibility is crucial. This is characterised on the one hand by the possibility of acquiring the assessments and on the other hand by the language in which they are presented, which represents an important barrier to access.

Language represents a fundamental dimension of accessibility that significantly influences the utilisation and reach of assessment tools. Atlas 2002 and Atlas Update were published exclusively in English, potentially limiting access for non-English-speaking evaluation communities. MESA studies are generally produced in English, and while some have been made available in additional languages depending on the country context, this has not been a systematic practice across all cases. While the Evaluation Globe volumes were initially published in English, the Americas compendium was subsequently translated into Spanish, demonstrating recognition of the diversity of linguistic requirements. INCE adopts the reverse approach, initially publishing all materials in Spanish before providing English translation; French translation is currently under way.

Publication format substantially affects accessibility across different user groups. Atlas 2002 was published in traditional book format, requiring purchase and thereby restricting access to individuals and institutions with adequate financial resources. Currently, despite its continuing relevance and usage, Atlas 2002 is out of print. Atlas Update is available free of charge as a downloadable publication, making this comprehensive resource accessible to anyone with internet connectivity. The three Evaluation Globe volumes that were already published at the time of writing are available only in book format, limiting accessibility despite their comprehensive global coverage. However, supplementary materials and selected aspects are available through Evaluation Globe's home page and international journal publications.

Contemporary assessment tools increasingly utilise digital platforms to enhance accessibility. All MESA studies are shared with the public once finalised. Out of the 13 MESA studies that have been conducted thus far, 6 are downloadable from the GEI website; these cover countries, institutions and a city across three continents.

The INCE data and materials are shared via the INCE home page, making questionnaires and results as well as comprehensive information on its development, the data collection tools and the measurement process freely available. INCE therefore offers a good basis for further research on NESs. However, the INCE home page requires adaptation to reflect its expanding global scope, as the current page primarily showcases the data for Latin America and the Caribbean. The maintenance of digital platforms, including data uploads, design updates and usability improvements, requires sustained resources that are not always available to coordinating institutions.

6.2 Global coverage of the assessments

The five assessment tools demonstrate varying approaches to global coverage, reflecting different strategic priorities and resource constraints. Atlas 2002 and Atlas Update focus primarily on OECD member states. In the original Atlas, 12 European, 2 North American, 3 Asian-Pacific and 1 African country are covered. The subsequent Atlas Update encompasses 12 European, 2 North American and 5 Asian-Pacific countries. This approach enables comparison among countries sharing key characteristics, such as income levels, democratic governance structures and education levels, thereby enhancing comparability, although significant developments in other regions are potentially overlooked. The key and defining criteria for country selection in Atlas 2002 was that the editors knew "that there were substantive and multiple evaluation activities in each country" (Furubo et al., 2002: 5); furthermore, they aimed to cover the globe as far as possible. While the editors would have liked to include more countries, they could not secure country chapters for them (Furubo et al., 2002).

Evaluation Globe addresses geographical limitations identified in earlier assessments. As stated in the preface, "the stark contrast between, on the one hand, our own experiences and the impressive dynamics of the institutionalisation of evaluation outside of North America and Europe and, on the other hand, the lack of attention given to research in these regions, gave rise to the idea that the venture to provide a comprehensive overview beyond our own borders was one worth attempting" (Stockmann et al., 2020: vi). Across its three volumes published at the time of writing, Evaluation Globe covers 38 countries and thus

represents the assessment with the highest diversity of countries examined (for direct comparison of countries included in the assessments, please see Annex 3).

MESA was developed in the context of international cooperation, as the GEI implements interventions in 19 countries in Africa, Asia and Latin America. Hence, it is not part of the GEI's mission or operational focus to conduct a MESA study in a European country. However, the instrument's free accessibility allows for potential implementation by interested experts in other countries or regions.

INCE currently covers 11 countries in Latin America and the Caribbean, 6 African countries and 1 Asian country, with planning under way for 4 additional African measurements. While INCE covers Spanish-speaking Latin America extensively, gaps remain regarding Portuguese- and English-speaking countries in the region and, perhaps more importantly, in other regions across the globe. INCE was developed as part of development cooperation and now concentrates on countries in Africa, Asia and Latin America. It is not expected that this will change in the near future unless countries outside of these regions are interested in measurements.

The measurement process demands participation from a variety of national actors, not all of whom understand Spanish, English or French. A translation of the INCE survey into the respective national language or widely used second language of a country is therefore a first step in measurement where English, French or Spanish are not official languages. This poses a challenge for INCE measurements worldwide, though one that can be overcome with the support of national evaluation associations, which often help with translation. This was the case in Mongolia, where the survey had to be translated into Mongolian. But this is also the case in most other countries where the INCE survey must be adapted to local contexts, as is common for surveys applied in multiple contexts (Behr, 2023)).

Certain assessment approaches have inherent coverage limitations. INCE measurements need the voluntary participation of national evaluation units, simultaneously representing both a strength and weakness. While this approach has the potential to generate ownership and avoid external imposition, it restricts implementation to countries with established evaluation units willing to invest the necessary time and resources. However, in countries where such units do exist and they assume full responsibility for the process, the additional financial cost of applying INCE can be minimal – virtually zero in some cases if the measurement is done by staff of the evaluation units and/or by volunteers from national evaluation associations. When external support is needed, costs remain relatively low, averaging around 3,000 US dollars per measurement and only increasing up to about 8,000 US dollars in contexts where an initial in-person workshop is deemed necessary to build capacity and foster collaboration (such as in some African countries). These relatively low costs, combined with the fact that the only formal requirement for application is the initiative and leadership commitment of the national evaluation unit, help explain the rapid expansion of INCE across Latin America and the Caribbean.

6.3 Summarizing analysis of accessibility and global coverage

Global coverage patterns reflect the historical development of evaluation systems across regions. Evaluation systems in North America began implementation in the 1960s, followed by Latin American systems approximately 10 years later, while African evaluation systems emerged roughly 20 years ago. The institutionalisation of evaluation in Asia commenced considerably later than in North America and developed heterogeneously across different countries. While systematic evaluation was already being conducted in India in the 1950s, many Asian countries remain in the initial stages of evaluation institutionalisation. The regional evaluation society, the Asia-Pacific Evaluation Association (APEA), was established in 2012 . European evaluation institutionalisation was significantly influenced by EU requirements (Martinaitis et al., 2019; Pattyn et al., 2017), while evaluation practice in many African, Asian and Latin American countries remains dominated by bilateral and multilateral international cooperation organisations.

Language accessibility patterns directly correlate with regional coverage strategies. INCE's initial development based on Spanish facilitated extensive Latin American coverage but created barriers for other regions. The subsequent English translation enabled African and Asian expansion, while ongoing French translation aims to enhance francophone accessibility. This pattern demonstrates how language decisions fundamentally shape the reach of assessment tools and regional penetration.

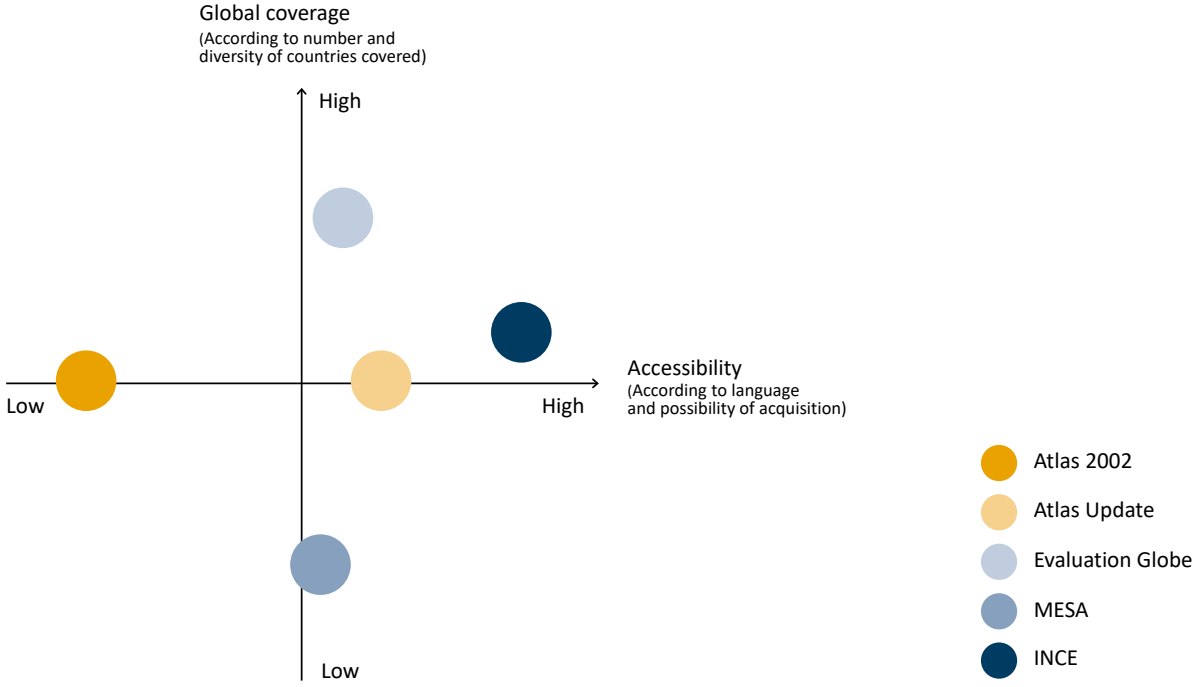
The relationship between assessment methodology and global coverage reveals systematic constraints. INCE's requirement for leadership of the measurements by the respective national evaluation unit, while generating ownership, inherently limits coverage to countries with established institutional structures. This methodological choice creates a paradox where the tool designed to strengthen evaluation systems cannot reach countries with the weakest institutional foundations. Similarly, MESA's focus on development cooperation contexts naturally concentrates on coverage in specific regions, excluding others.

While online platforms theoretically enable universal access, in practice, accessibility depends on internet infrastructure, digital literacy and ongoing platform maintenance. The need for adaptation of the INCE home page to ensure global scope illustrates how digital accessibility requires sustained resource investment beyond initial development. This highlights the challenge of maintaining equitable access across regions with varying technological infrastructure.

The comparative analysis in this discussion paper reveals that achieving both comprehensive accessibility and extensive global coverage requires substantial ongoing resource commitment. Free access models, while democratizing access, depend on sustained funding for platform maintenance, translation services and content updates. This creates sustainability challenges that may ultimately constrain both accessibility and expansion of coverage over time.

The assessment tools demonstrate an evolutionary pattern from geographically concentrated, academically oriented resources towards more accessible, globally distributed diagnostic tools (see Figure 8). This simultaneously reflects a general and pivotal development in international development cooperation: decision-making and implementation are increasingly shifting away from international experts (predominantly from the Global North) towards experts from the countries where the cooperation takes place. However, this evolution reveals persistent tensions between comprehensive coverage and practical accessibility. Future development of assessments must address these fundamental trade-offs through innovative approaches that balance methodological rigor, global representation and equitable access while also ensuring long-term sustainability.

Figure 8 Classification of the five assessments according to global coverage and accessibility[6]



Source: DEval, own visualisation

[6] The classification relates only to the published assessments. The analysis carried out in this discussion paper forms the basis for the classification of the individual assessments. The table in Annex 5 provides more detailed information on the assessment of the individual categories of global coverage and accessibility.

7. CONCLUSION AND OUTLOOK

The assessments described in this discussion paper present valuable inputs for the ongoing discussion and process of learning about NESs. The diversity and the continuous development of NESs pose a challenge to standardised measurement and comparison of assessments, but all of the presented works demonstrate that this endeavour is possible. The fact that Evaluation Globe, MESA and INCE were developed and applied over the last few years show the great interest in NESs on a global scale. This gives reason to assume that the coming years will see more formalised and strengthened NESs. Four of the described assessments were also applied to evaluation systems in organisations. INCE has not been used for this purpose, but piloting it with an organisation could add further value to the instrument.

Atlas 2002 and Atlas Update laid the foundation for the analysis of NESs. This pioneering work made it possible to carry out further systematic analyses of NESs. The Atlas approach provides the basis for more recent work in this area, and the findings from Atlas 2002 are still relevant and much used.

Evaluation Globe develops Atlas further, not only in that it can actually apply to the "globe" geographically and depicts almost all regions of the world, but also because it offers much deeper analyses, including extremely valuable regional comparative analyses. The network that has developed for the creation of Evaluation Globe alone should be very important for further work on the subject. Today, Evaluation Globe offers an invaluable treasure trove of data and analyses on NESs, and the challenge will be to keep it up to date. A link with INCE could be useful here to a limited extent. The INCE measurements are ideally repeated every two years, but INCE does not cover the same countries as Evaluation Globe and will not be able to do so in the foreseeable future due to the limitations described above. In addition, the methodological approaches of the two instruments differ significantly. However, it is certainly conceivable that a second edition of the Evaluation Globe volumes could use existing INCE data and that future INCE-Briefs could use the Evaluation Globe data. A further possibility lies in the linkage between the data collection processes: as part of the INCE measurement, relevant institutions often reflect on their answers, and this reflection process could allow for further data collection covering the Evaluation Globe questions. A similar process could create a stronger linkage between MESA and INCE.

MESA studies already use INCE data, where available. In general, MESA offers an excellent way to add depth to the more superficial INCE analysis. While INCE is useful for identifying which components of an evaluation system are strong or weak, it does not always explain the underlying causes or key factors behind this. Many stakeholders prefer a comprehensive report to the brief INCE country reports with graphical presentation of data. Stakeholders want to know more precisely which components of the system are working well, which should be strengthened further and why, and this is exactly what MESA offers. So far, it has only been possible to produce a MESA study following an INCE measurement on two occasions – in Colombia and Guatemala. However, the MESA studies were able to go into depth in precisely the areas that were identified as weak by the INCE measurement. It is to be hoped that this intermeshing of instruments will happen more often in the future, which is a desire of all actors involved and is being made possible step by step.

The freely available INCE data is of great value to scientists and evaluation (capacity development) practitioners. To date, INCE data has focused mainly on one region and has therefore been of little interest to an international audience. It might be that measurements can be carried out in more regions and countries and this valuable data used for more in-depth analysis, allowing for further learning on NESs. Analysis of INCE data has so far been rather superficial, but it would be a good source of data to examine potential correlations within evaluation systems, such as the relationship between the two INCE dimensions of institutionalisation and use of evaluation (similar to analysis undertaken in Evaluation Globe). INCE data can be used in very practical terms, feeding directly into plans for strengthening evaluation systems. However, INCE data could also become more useful for academics interested in examining influencing factors for strengthening NESs. For instance, correlations between individual variables could be investigated, or trends across countries could be analysed.

For ECD practitioners, analysing their own work based on data from the instruments presented here would of course be particularly valuable. For example, the INCE measurement in Mongolia serves as a baseline study for the ECD work by DEval since 2025. The Evaluation Globe country studies can play a similar role. However, since neither INCE nor Evaluation Globe were developed based on ECD interventions, they cannot provide data for a robust link between ECD interventions and the strengthening of NESs, though they can certainly provide data for important correlational analyses.

The majority of ECD interventions today are trainings. Only a few actors carry out further work, such as learning-by-doing evaluations, strengthening the institutionalisation of evaluation, or the establishment of national evaluation networks or platforms. It would be desirable if data on evaluation systems were better used in the future to analyse ECD approaches (training alone vs. systemic approaches) in order to learn more about which instruments are particularly helpful for strengthening NESs. This data would certainly come to a large extent from the Global South and could be used to build and strengthen NESs in the Global North.

According to Mackay (2007: 68): “In this environment, it is important to regularly monitor and evaluate the M&E system itself—just as any area of public sector reform should be regularly assessed. Indeed, conducting regular M&E efforts to strengthen an M&E system is one way those in charge of such efforts can lead by example.” INCE answers this demand by regularly assessing NESs.

However, Dahler-Larsen and Boodhoo (2019: 278) note: “It is a massive undertaking to collect valid and reliable comparable data about national evaluation cultures.” And INCE shows that this is correct. INCE is indeed a substantial undertaking, requiring considerable coordination efforts and voluntary input from diverse actors. As described above, this joint work between a diverse set of actors is a strength of INCE, which fosters a vibrant community of practice and allows for learning on both evaluation systems and harmonisation of capacity development. In fact, INCE’s greatest added value lies in part in the networks of collaboration it generates. At the same time, this may be a weak point, because the instrument can only perform well when those who manage and use it maintain interest. In other words, INCE can only be maintained if it keeps its relevance. The use of INCE data may motivate and inspire more countries to apply INCE, and thus measurements and use of the data will hopefully reinforce each other.

8. REFERENCES

Behr, D. (2023), "What to Consider and Look Out For in Questionnaire Translation", GESIS – Leibniz Institute for the Social Sciences, Mannheim.

von Bertalanffy, K. L. (1968), *General System Theory: Foundations, Development, Applications, Rev. Ed.*, Braziller, New York.

Bohr, N. (1928), "The Quantum Postulate and the Recent Development of Atomic Theory", *Nature*, Vol. 121, pp. 580–590.

Bustelo, M. and S. Jacob (2025), "From Studies to Systems: Ray Rist's Influence on Evaluation Systems: Insights from International Research Group for Policy and Program Evaluation (INTEVAL)", *Journal of MultiDisciplinary Evaluation*, Vol. 21, No. 50, pp. 123–129.

Chirau, Takunda et al. (2020), "A Stakeholder View of the Development of National Evaluation Sysems in Africa.", *African Evaluation Journal*, Vol. 8, No. 1, p. a504.

Dahler-Larsen, P. and A. Boodhoo (2019), "Evaluation Culture and Good Governance: Is There a Link?", *Evaluation*, Vol. 25, p. 135638901881911.

Dahler-Larsen, P. and E. Raimondo (2024), "The Sceptical Turn in Evaluation and What to Do with It: Keynote Presentation Delivered by Peter Dahler-Larsen and Estelle Raimondo at the EES Conference in Copenhagen, June 10, 2022", *Evaluation*, Vol. 30, No. 1, pp. 69–81.

Departamento Nacional de Planeación and Consejo Nacional de Política Económica y Social (2022), "Fortalecimiento del uso y la institucionalidad de las evaluaciones para la toma de decisiones en Colombia", Regierungsdokument, No. CONPES 4083, Bogotá.

Diwakar, Y. and B. Rao (2023), "Evolution of the National Evaluation System in India", Vol. 1, pp. 47–73.

Diwakar, Yatin et al. (2022), "A Study on the Status of National Evaluation Policies and Systems in the Asia Pacific Region", Asia Pacific Evaluation Association.

Eckhard, S. et al. (2024), "Institutional Design and Biases in Evaluation Reports by International Organizations", *Public Administration Review*, John Wiley & Sons, Ltd, Vol. 84, No. 3, pp. 560–573.

Feinstein, O. (2015), "On the Development of Evaluation Systems in Latin America and the Caribbean", *Revista Del CLAD Reforma y Democracia*, pp. 193–210.

Feinstein, O. N. (2014), "La Institucionalización de La Evaluación de Políticas Públicas En América Latina", *Presupuesto y Gasto Público*, Vol. 68, pp. 14–52.

Furubo, J.-E. et al. (Eds.) (2002), *International Atlas of Evaluation*, Transaction Publishers, New Brunswick.

Gaarder, M. and B. Briceno (2010), "Institutionalisation of Government Evaluation: Balancing Trade-Offs", *The Journal of Development Effectiveness*, Vol. 2, pp. 289–309.

Global Evaluation Initiative (2022), "MESA Guidance Notes".

Global Evaluation Initiative (**no date**), "MESA: GEI's Diagnostic Tool for a Monitoring and Evaluation Systems Analysis", *Global Evaluation Initiative*, *https://www.globalevaluationinitiative.org/mesa*.

Goldman, I. et al. (2018), "The Emergence of Government Evaluation Systems in Africa: The Case of Benin, Uganda and South Africa", *African Evaluation Journal*, Vol. 6, No. 1.

Head, B. W. (2015), "Toward More "Evidence-Informed" Policy Making?", *Public Administration Review*, Vol. 76, No. 3, pp. 472–484.

Hollstein, B. (2021), "Promises and Pitfalls of Qualitative Longitudinal Research", *Longitudinal and Life Course Studies*, Bristol University Press, Vol. 12, No. 1, pp. 7–17.

Hummelbrunner, R. (2011), "Systems Thinking and Evaluation.", *Evaluation*, Vol. 17, No. 4, pp. 395–403.

Jacob, S. (2023), "The Institutionalization of Evaluation around the Globe: Understanding the Main Drivers and Effects over the Past Decades", *Handbook of Public Policy Evaluation*, Edward Elgar Publishing Limited, Northampton, 1st ed., pp. 187–205.

Jacob, S. et al. (2015), "The Institutionalization of Evaluation Matters: Updating the International Atlas of Evaluation 10 Years Later", *Evaluation*, Vol. 21, No. 1, pp. 6–31.

Kalugampititya, A. R. L. (2022), "Critical Factors for Institutionalizing Evaluation at National Level: Study on Four Countries in Asia - Sri Lanka, Nepal, Philippines and Bangladesh", Universität des Saarlandes, Saarbrücken, Germany.

Keating, C. B. et al. (2020), "Systems Theory: Bridging the Gap Between Science and Practice for Systems Engineering.", *INCOSE International Symposium*, Vol. 30, No. 1, pp. 1017–1031.

Klier, S. et al. (2022), "Grounding Evaluation Capacity Development in Systems Theory", *Evaluation*, Vol. 28, No. 2, pp. 231–251.

Lahey, Robert (2025), "The Road Toward Institutionalizing Evaluation in Developing Countries: Following the Path of Ray Rist", *Journal of MultiDisciplinary Evaluation*, Vol. 21, No. 50, pp. 104–110.

Leeuw, F. and J.-E. Furubo (2008), "Evaluation Systems What Are They and Why Study Them?", *Evaluation*, Vol. 14, pp. 157–169.

Mackay, K. (2007), "How to Build M&E Systems to Support Better Government", No. 40546, World Bank Group, Washington D.C, USA.

Mackay, K. (2009), "Building Monitoring and Evaluation Systems to Improve Government Performance.", *Country-Led Monitoring and Evaluation Systems. Better Evidence, Better Policies, Better Development Results.*, Geneva, Switzerland, pp. 169–187.

Martinaitis, Ž. et al. (2019), "Evaluation Systems: How Do They Frame, Generate and Use Evidence?", *Evaluation*, Vol. 25, No. 1, pp. 46–61.

Merton, R. K. (1940), "Bureaucratic Structure and Personality", *Social Forces*, Vol. 18, No. 4, pp. 560–580.

Meyer, W. and R. Stockmann (2020), "Institutionalisierung Der Evaluation in Den Politischen Systemen Europas. Eine Vergleichende Analyse", *Dms - Der Moderne Staat - Zeitschrift Für Public Policy, Recht Und Management*, Vol. 1, pp. 24–43.

Nardo, M. et al. (2008), "Handbook on Constructing Composite Indicators: Methodology and User Guide", OECD Publishing.

OECD (2020), *Improving Governance with Policy Evaluation: Lessons from Country Experiences.*, OECD Public Governance Reviews, OECD Publishing, Paris.

Pattyn, V. and M. Bouterse (2020), "Explaining Use and Non-Use of Policy Evaluations in a Mature Evaluation Setting", *Humanit Soc Sci Commun*, Vol. 85, No. 7, doi: *https://doi.org/10.1057/s41599-020-00575-y*.

Pattyn, V. et al. (2017), "Policy Evaluation in Europe", *The Palgrave Handbook of Public Administration and Management in Europe*, pp. 577–593.

Pérez Yarahuán, G. (2015), *Panorama de Sistemas Nacionales de Monitoreo y Evaluación En América Latina*.

Porter, S. and I. Goldman (2013), "A Growing Demand for Monitoring and Evaluation in Africa", *African Evaluation Journal*, Vol. 1, No. 1, p. 9.

Raimondo, E. (2018), "The Power and Dysfunctions of Evaluation Systems in International Organizations", *Evaluation*, Vol. 24, pp. 26–41.

Rosenstein, B. (2013), "Mapping the Status of National Evaluation Policies. Parliamentarians Forum on Development Evaluation in South Asia and EvalPartners."

Rosenstein, B. (2015), "Mapping the Status of National Evaluation Policies, 2nd. Edition", Parliamentarians Forum on Development Evaluation in South Asia, EvalPartners.

Rosenstein, B. and A. Kalugampitiya (2021), "Mapping of the Status of NEPs 2021", EvalPartners; Global Parliamentarians Forum for Evaluation (GPFE), n.p.

Saltelli, A. (2007), "Composite Indicators Between Analysis and Advocacy", *Social Indicators Research*, Vol. 81, No. 1.

Stockmann, R. et al. (Eds.) (2022a), *The Institutionalization of Evaluation in the Americas*, Palgrave MacMillan, Cham, Switzerland.

Stockmann, R. et al. (Eds.) (2020), *The Institutionalization of Evaluation in Europe*, Palgrave MacMillan, Cham, Switzerland.

Stockmann, R. et al. (2022b), *Evaluation GLOBE - Compendium on the Institutionalization of Evaluation. Volume IV: Africa.*, Concept Paper for the IV Volume, Saarbrücken.

Stockmann, R. et al. (Eds.) (2024), *The Institutionalization of Evaluation in Asia-Pacific*, Palgrave MacMillan, Cham, Switzerland.

Toulemonde, J. (2000), "Evaluation Culture(s) in Europe: Differences and Convergence Between National Practices", *Vierteljahrshefte Zur Wirtschaftsforschung / Quarterly Journal of Economic Research*, Vol. 69, pp. 350–357.

Weber, M. (1922), *Wirtschaft Und Gesellschaft: Grundriss Der Verstehenden Soziologie*, J.C.B. Mohr (Paul Siebeck), Tübingen.

Widmer, T. et al. (Eds.) (2009), *Evaluation: Ein Systematisches Handbuch*, Springer, Wiesbaden.

Williams, B. and I. Imam (Eds.) (2006), *Systems Concepts in Evaluation: An Expert Anthology*, American Evaluation Association, Point Reyes, CA, USA.

9. ANNEX

Annex 1 Actors involved in the initial development of INCE

International organisations & UN agencies	Regional & bilateral organisations	National public institutions & ministries	Civil society organisations/ NGOs	Academic institutions
- WFP - CLEAR-LAC - UNICEF - UN Women - United Nations Population Fund (UNFPA) - IDB	- DEval - Latin American and Caribbean Network for Monitoring, Evaluation and Systematisation (Red de Seguimiento Evaluación y Sistematización de América Latina y el Caribe, ReLAC) - RIEPP (Red Internacional de Evaluación de Políticas Públicas)	- CONEVAL (Mexico) - Evalúa Jalisco (Mexico) - Social Policy Cabinet (Gabinete de Coordinación de la Política Social, GCPS)/Ministry of Economy, Planning, and Development (Ministerio de Economía, Planificación y Desarrollo, MEPyD) (Dominican Republic) - MIDEPLAN (Costa Rica) - Ministry of Development and Social Inclusion (Ministerio de Desarrollo e Inclusión Social, MIDIS)/National Centre of Strategic Planning (Centro Nacional de Planeamiento Estratégico, CEPLAN) (Peru) - National Planning Secretariat (Secretaría Nacional de Planificación, SNP) (Ecuador) - System of Information, Evaluation and Monitoring of Social Programs (Sistema de Información, Evaluación y Monitoreo de Programas Sociales, SIEMPRO) (Argentina) - National Planning Department (Departamento Nacional de Planeación, DNP) (Colombia) - Secretariat of Planning and Programming of the Presidency (Secretaría de Planificación y Programación de la Presidencia, SEGEPLAN) (Guatemala)	- Grupo Faro (Ecuador) - TECHO	- University of Chile (Universidad de Chile) - National University of San Juan Argentina (Universidad Nacional de San Juan) - Technical University of Uruguay (Universidad Tecnológica del Uruguay)

Annex 2 Purpose of the assessments

	Atlas 2002	Atlas Update	MESA	Evaluation Globe	INCE
Direct contribution to strengthening NESs	Primarily describes evaluation system situations and developments, serving as reference material rather than providing actionable ECD frameworks	Focuses on analysing trends and institutionalisation patterns useful for understanding context, but lacks direct operational guidance for systematic capacity development interventions	Explicitly designed to "inform capacity-development strategies meant to strengthen" M&E systems and serves "as a basis for a national ECD strategy"; not an end in itself but specifically structured to guide ECD planning	Explicitly aimed at "evaluation capacity development" and "serves as a basis for presenting particularly promising components of national evaluation systems", but requires more interpretation to develop specific ECD strategies	Systematically designed to "enhance national evaluation systems" through peer-to-peer learning and network development; built-in collaborative learning paradigm directly supports capacity-building objectives with regular monitoring for dynamic capacity evolution
Value for scholarly discussion	Developed by evaluation scholars with "extensive academic credentials in evaluation theory"; published as an academic book with systematic country-by-country analysis and established as "seminal work" widely utilised by scholars and practitioners	Scientific article with longitudinal analysis methodology; builds on established theoretical framework and provides systematic comparative analysis with clear research objectives across 19 OECD countries	Described as a "diagnostic tool" with flexible implementation; focused on practical application and ECD planning, with limited description of explicit academic methodology or theoretical framework in the available information	Developed by academics and evaluation experts with a comprehensive methodology involving 100+ evaluation experts worldwide, but more practice-oriented, aimed explicitly at an interdisciplinary audience, including politicians and administrative staff	Developed through a systematic participatory process with diverse stakeholders; includes regular measurement cycles and systematic methodology, but designed primarily for practical capacity development rather than academic research

Annex 3 List of countries analysed in the five assessments

Atlas 2002	Atlas Update	MESA	Evaluation Globe	INCE
Australia	Australia	Brazil	Argentina	Argentina
Canada	Canada	Colombia	Australia	Benin
China	Denmark	Dominican Republic	Bangladesh	Chile
Denmark	Finland	Jamaica	Belgium	Colombia
Finland	France	Madagascar	Bolivia	Congo-Brazzaville
France	Germany	Mozambique	Brazil	Costa Rica
Germany	Ireland	Saint Lucia	Canada	Dominican Republic
Ireland	Israel	Solomon Islands	Chile	Ecuador
Israel	Italy	Uzbekistan	China	Gabon
Italy	Japan		Colombia	Ghana
Japan	Netherlands		Costa Rica	Guatemala
Korea	New Zealand		Czechia	Ivory Coast
Netherlands	Norway		Denmark	Mexico
New Zealand	South Korea		Ecuador	Mongolia
Norway	Spain		Finland	Morocco
Spain	Sweden		France	Paraguay
Sweden	Switzerland		Germany	Peru
Switzerland	United Kingdom		India	South Africa
United Kingdom	United States		Ireland	Sri Lanka
United States			Italy	Tanzania
Zimbabwe			Japan	Uganda
			Latvia	Uruguay
			Mexico	Zimbabwe
			Nepal	
			Netherlands	
			Pakistan	
			Peru	
			Philippines	
			Poland	
			Portugal	
			Romania	
			South Korea	
			Spain	
			Sri Lanka	
			Switzerland	
			Taiwan	
			United Kingdom	
			United States	
21	**19**	**9**	**38**	**23**

Annex 4 Level of leadership of the assessments undertaken by national stakeholders in the countries assessed

	Atlas 2002	Atlas Update	MESA	Evaluation Globe	INCE
Development	**Low** – This was developed mainly by Northern scholars; national experts contributed country chapters but did not contribute to framework design.	**Low** – The framework was built on Atlas 2002, led by academics in the North.	**Medium** – This was designed by GEI with practitioner input; the ownership principle is explicit, but not in the original design.	**Medium** – The framework was set by Global North academics, but chapters were authored by 100+ national experts worldwide.	**High** – This was co-created by Latin American governments, evaluation units, VOPEs and multilaterals; the design is participatory.
Application	**Medium** – Country experts collected and analysed data for their chapters, but under external editorial guidance.	**Medium** – Again, country experts wrote chapters, but the process was less participatory than in MESA/INCE.	**High** – This is co-led by national institutions; country demand is required before implementation.	**Medium** – National scholars contributed data, ensuring contextual detail, but the process was not always government-led.	**High** – This is implemented by national evaluation units; there is high stakeholder participation via surveys and workshops.
Use	**Low** – Results are published in a book; there is little evidence of governments using them for their own strategies.	**Low** – This is mainly for academic use; there is limited evidence of uptake by governments.	**High** – Results feed directly into ECD strategies, often shaping government actions.	**High** – Countries integrate INCE into national evaluation policies, strategies and development plans (e.g. in Costa Rica, Colombia and Peru).	**High** – Countries integrate INCE into national evaluation policies, strategies and development plans (e.g. in Costa Rica, Colombia and Peru).

Annex 5 Global coverage and accessibility of published assessments

	Atlas 2002	Atlas Update	MESA	Evaluation Globe	INCE
Global coverage of available assessments	- 21 countries (12 European, 2 North American, 6 Asian-Pacific, 1 African) - Focus on OECD member states	- 19 countries (12 European, 2 North American, 5 Asian-Pacific) - Focus on OECD member states	- 5 MESA country studies across three continents are available online (out of 13 MESA studies developed so far) - Focus on Africa, Asia and Latin America - Continuous expansion; more MESA studies have been developed but not yet shared with the wider public	- 38 countries across three continents - Covers Africa, Asia and Latin America - Multi-continent assessment	- 19 countries (11 Latin American/Caribbean, 6 African, 1 Asian), - Dependent on national evaluation units
Accessibility	- English only - Book form – purchase required - Book is out of print	- English only - Free online download	- Studies published in the official language of the respective country/organisation - Only selected studies are available for free download from the GEI website	- English; Americas volume translated into Spanish - Book form – purchase required	- Data/information available in Spanish, English and French - Available for free via the INCE home page