

1.1

Conceptualizing National M&E Systems: A Scoping Review

Douglas Glandon
Guilherme Cunha Lima

Abstract

This scoping review traces the evolving conceptualization of Monitoring and Evaluation (M&E) (eco) systems over the past four decades, from a technical tool for donor accountability to a dynamic ecosystem central to learning and systemic change. This synthesis follows a scoping-review protocol grounded in JBI guidance, with reporting per PRISMA-ScR. It examines how M&E has shifted across three historical phases—from basic data collection in the 1980s and 1990s, through Results-Based Management in the 2000s, to holistic, context-sensitive frameworks in the 2010s and beyond. This evolution reflects an expansion beyond compliance and project oversight toward adaptive, learning-oriented processes that incorporate stakeholder engagement, institutional culture, and local ownership. The review also synthesizes prominent conceptual frameworks, tracing a progression from early technical models centered on measurement and accountability to those emphasizing leadership, values, relational dynamics, and political will. Recent frameworks underscore the importance of enabling environments and collaborative networks, highlighting the interdependency of technical capacity, governance structures, and cultural contexts. Overall, M&E is increasingly conceptualized as a system of interlinked actors and processes, shaped by changing development paradigms and societal expectations. While foundational models remain influential, future conceptualizations will likely incorporate emerging priorities into more flexible, responsive frameworks that balance accountability with inclusivity and ongoing innovation.

Introduction

The conceptualization of Monitoring and Evaluation (M&E) (eco)systems has evolved significantly over the last four decades, mirroring broader shifts in development paradigms, governance priorities, and societal expectations. This scoping review seeks to answer the research question: How are M&E (eco)systems conceptualized in the literature? It explores this evolution, tracing M&E from its origins as a technical tool for accountability to its current framing as a dynamic, context-sensitive ecosystem integral to learning, decision-making, and systemic change. The review reveals a progressive expansion in scope – moving beyond mere data collection to encompass institutional frameworks, cultural influences, stakeholder engagement, and adaptive processes – reflecting M&E's growing complexity and relevance in contemporary governance. To organize these insights, the review is structured into three main sections. First, A Brief (Recent) History of M&E System Conceptualization traces the field's development across three distinct periods: the 1980s–1990s, when M&E served as a technical instrument for donor accountability; the 2000s, defined by the rise of Results-Based Management (RBM) and the expansion of national M&E systems in developing countries; and the 2010s–present, characterized by an integrated, ecosystem model prioritizing learning and stakeholder engagement. Second, Conceptual Frameworks for M&E Systems analyzes the theoretical models underpinning these conceptualizations, progressing from early technical frameworks to recent approaches that integrate political, relational, and cultural dimensions. Finally, the conclusion synthesizes the findings, reflects on the trajectory of M&E system conceptualization, and considers directions for future research.

Conceptual Framework

This scoping review is one of a set of twenty scoping reviews commissioned to inform the Global Evaluation Initiative (GEI) 2.0 Strategy (FY2026–30).

The review set is organized around a conceptual framework that mirrors the nested structure of GEI's theory of change and theory of action which addresses four analytical levels: country, decision-maker/unit, evaluation/decision, and individual. At the country level, the focus is on the linkage between stronger M&E ecosystems and development and governance outcomes. At the decision-maker/unit

level, the emphasis is on strengthening the culture and practice of evidence-informed decision-making within specific ministries or units. At the evaluation/decision level, the focus is on how evaluations or other evidence-generating activities can inform specific policy or program decisions. Finally, at the individual level, the framework addresses how training, mentorship, peer exchange, and related supports strengthen the skills, performance, and professional development of evaluators and other evaluation actors. Within each level, the review set addresses five recurring sub-topics: how the focal construct is defined and conceptualized; the primary theories of change linking the level's activities to its intended outcomes; the metrics used to characterize status and change at that level; the empirical evidence linking activities to outcomes; and the drivers that help explain success or failure. The resulting 4×5 matrix — four analytical levels by five recurring sub-topics — structures the review set into twenty distinct but complementary scoping reviews.

The present review addresses the *definition and conceptualization* sub-topic at the *country* level.

Methods

Review design

This review was conducted as a scoping review following JBI guidance (Peters et al., 2020), with reporting per PRISMA-ScR (Tricco et al., 2018). A scoping format was appropriate because the research question is definitional and required breadth across heterogeneous, multidisciplinary material — peer-reviewed scholarship, institutional reports, and policy documents. Eligibility followed the JBI Population–Concept–Context (PCC) framing, with the review question and inclusion criteria anchored primarily to the concept of country-level M&E (eco)systems. Per §11.1.1 of the JBI Manual, no formal critical appraisal or risk-of-bias assessment was applied.

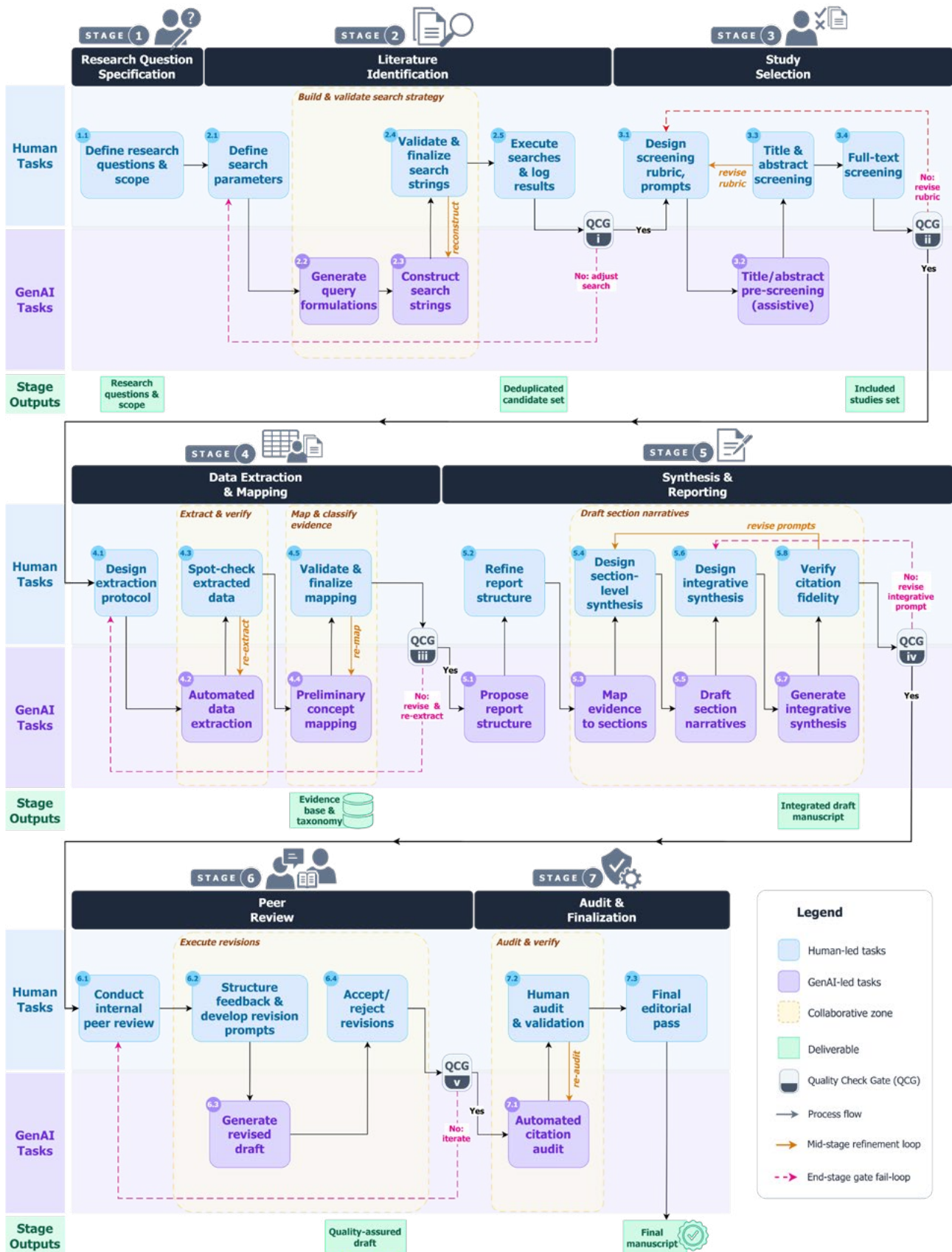
The review followed the seven-stage human-in-the-loop workflow shown in **Figure 1**, beginning with human-led specification of the research question and scope (Stage 1, Task 1.1). Across all stages, LLM tools supported defined tasks under human-designed prompts; two human reviewers retained authority over every inclusion, charting, mapping, and synthesis decision. Figure 1's five "collaborative zones" mark groupings of paired human and GenAI tasks that operated on the same evolving artifact through iterative cycles, with each actor's work shaping the other's next task.

Eligibility criteria

Sources were eligible if they articulated or analyzed how country-level M&E systems or ecosystems have been conceptualized or defined, in the context of national, whole-of-government, country-led, or national evaluation system settings. Eligible source types included peer-reviewed journal articles and book chapters; institutional reports and guidance documents from multilateral and bilateral development organizations and regional evaluation partnerships; government policy and evaluation-system documents; and doctoral theses directly addressing M&E system conceptualization. No date restriction was applied, because tracing the historical evolution of the construct is integral to the question. English-language sources were the primary focus. At full-text review, sources were excluded for failing to engage with definitional or conceptual aspects of country M&E ecosystems, invoking the term without defining or operationalizing it, falling outside the review's scope, or not meeting the minimum quality threshold applied to their source type.

Two source-type-specific quality thresholds were applied. Peer-reviewed sources had to have been issued through standard editorial peer-review processes. Grey-literature sources additionally had to meet three criteria: (i) offer an explicit and defended definition or conceptualization of the construct, rather than invoking the terminology in passing; (ii) a documented evidence or reference base; and (iii) identifiable institutional quality control (verified pragmatically by checking whether the source was published through an official institutional channel and bore clear markers of institutional authorship, approval, or accountability).

Figure 1. GEI 2.0 Rapid Evidence Review (RER) Workflow



Information sources and search strategy

Literature identification (Stage 2 of Figure 1) began with the human reviewer's definition of search parameters (Task 2.1). Scopus and Web of Science Core Collection provided primary peer-reviewed coverage, with Google Scholar as a complementary academic search engine and for cited-reference chaining. Four institutional registers and repositories were systematically searched for development-sector grey literature: the World Bank Open Knowledge Repository, the UNEG Publications Repository, the OECD iLibrary, and Global Evaluation Initiative (GEI) network publications. Coverage was extended through backward and forward citation chasing, expert recommendation, and key-author tracking focused on scholars whose prior work has shaped the conceptualization of country-level M&E ecosystems.

Within the "build and validate search strategy" collaborative zone, the LLM generated semantically equivalent query reformulations and drafted database- and repository-specific search strings (Tasks 2.2–2.3), which the human reviewer audited and finalized (Task 2.4); a search log archived run dates, applied limits, and result counts. Searches were executed under a pre-specified stopping rule (Task 2.5): searching continued through successive results pages until three consecutive pages yielded no further relevant records. **Quality Check Gate (QCG) i** required the deduplicated candidate set to meet coverage and retrievability criteria; where it did not, reviewers revised search parameters or strings and re-executed searches.

Source selection

Study selection (Stage 3) operated under a human-designed screening protocol (Task 3.1). A total of 385 records were identified through databases and registers (356 from databases; 29 from registers), with a further 107 identified through backward and forward citation searching, expert recommendation and known-source inclusion, and key-author tracking (**Figure 2**). After deduplication and metadata-based screening to remove document types outside the eligible scope — such as conference abstracts, editorials, and book reviews — 290 records underwent title and abstract screening: the LLM produced assistive inclusion-likelihood labels (Task 3.2), and two independent human reviewers performed the authoritative screening (Task 3.3), informed but not bound by the LLM labels. Forty-seven records advanced to full-text screening (Task 3.4); four could not be retrieved, leaving 43 assessed, of which 23 were excluded with coded reasons. In parallel, 104 of the 107 supplementary-stream records were retrieved and assessed (three were not retrievable), and 90 were excluded with coded reasons. A total of 34 sources were included in the review.

QCG ii required the included studies set to satisfy protocol-specified eligibility coverage; where it did not, the reviewers revised the screening protocol and re-screened the affected records.

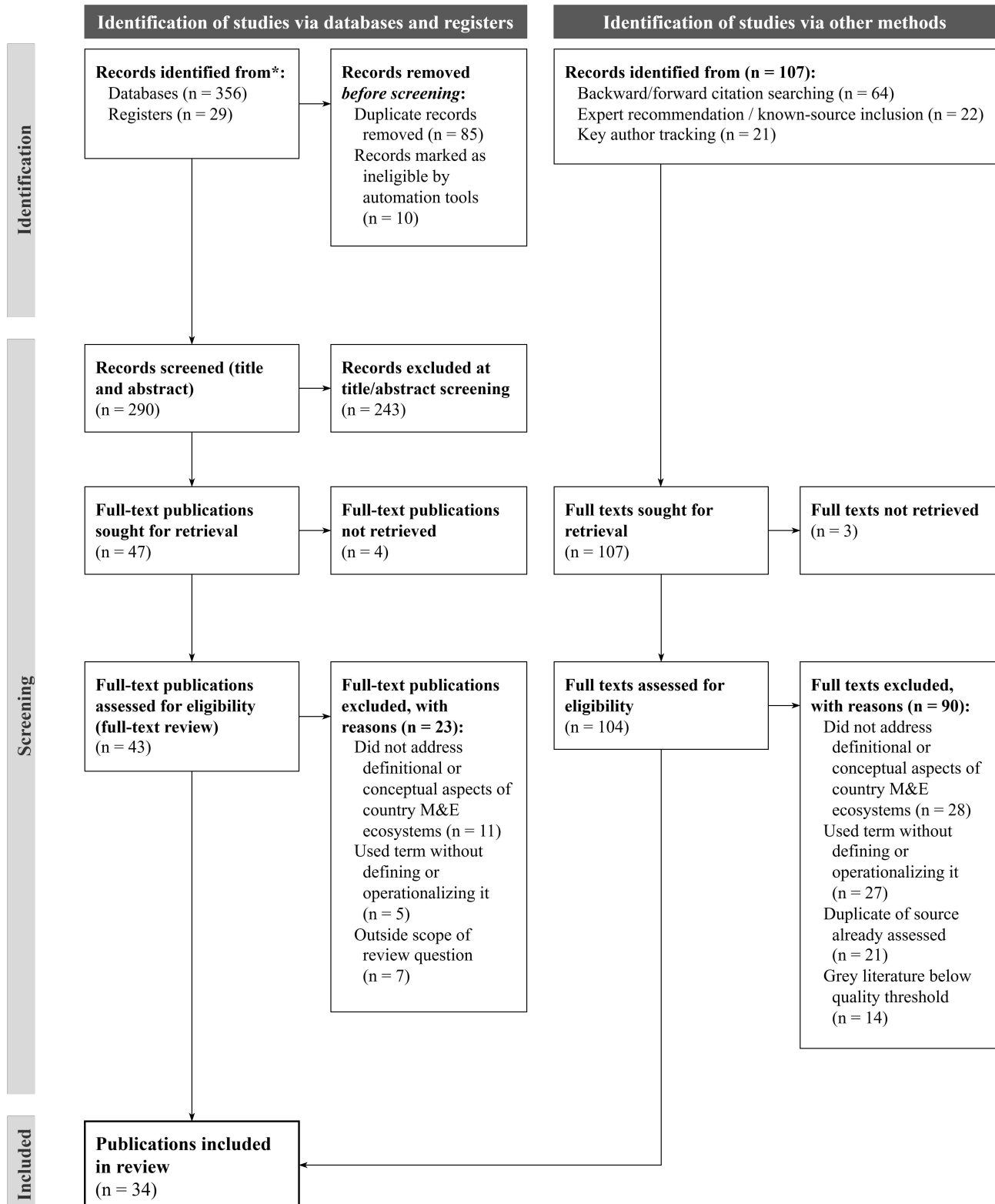
Data extraction and mapping

Data extraction and mapping (Stage 4) proceeded under a human-designed extraction protocol (Task 4.1). A structured codebook captured bibliographic metadata; the historical period addressed; the principal definition or conceptualization offered; the conceptual lineage invoked (e.g., results-based management, country-led evaluation); and the framework advanced.

Within the "extract and verify" collaborative zone, LLM-assisted extraction (Task 4.2) populated the codebook with page-anchored quotations and paraphrases, with iterative spot-checks and prompt refinement (Task 4.3) used to re-extract fields where verification flagged errors — including misattributions, unsupported statements, hallucinated content, and conflation of proximate constructs. Reviewers verified all definitional statements and framework attributions and a random 20% sample of remaining fields. Within the "map and classify evidence" zone, a preliminary concept mapping produced under structured prompts (Task 4.4) was then validated and finalized by the reviewers (Task 4.5), clustering the included sources along two axes: the historical era they addressed (1980s–1990s; 2000s; 2010s–present) and the conceptual framework they advanced.

QCG iii required the resulting evidence base and taxonomy to meet extraction-fidelity criteria — including page-anchored support for every charted statement — before synthesis. Where these criteria were not met, the process looped back to Task 4.1, where the reviewers revised the extraction protocol before the affected fields were re-extracted.

Figure 2. PRISMA Flow Diagram



Synthesis and reporting

Synthesis and reporting (Stage 5) proceeded through collaborative drafting cycles. A draft report structure was proposed under structured prompts (Task 5.1) and refined by the reviewers (Task 5.2). Within the "draft section narratives" collaborative zone, evidence was mapped to sections (Task 5.3), section-level synthesis logic was designed under human direction (Task 5.4), and section narratives were generated under human-designed prompts (Task 5.5) — producing, in this review, a two-part analytical structure spanning the historical evolution of the construct and the principal frameworks that have been advanced. An integrative prompt designed by the reviewers (Task 5.6) was then used to produce the integrated draft (Task 5.7).

A human verification of citation fidelity (Task 5.8) sampled a random 20% of cited claims and checked each against its source. Where errors, misattributions, or hallucinations appeared in the sample, the reviewer examined whether they were scattered across sections and sources or concentrated in one section or source; based on that assessment, the process looped back to Task 5.4, where the section-level synthesis prompts were revised and the affected section narratives regenerated. Where the sample was clean, the process advanced to QCG iv.

QCG iv assessed the structural and rhetorical quality of the integrated draft — whether the sections were well tied together, built on one another, and formed a coherent whole with clear flow and consistent framing. Where it did not, the reviewers revised the integrative synthesis prompt (Task 5.6) and the LLM regenerated the integrated draft.

Quality assurance

Before finalization, the integrated draft was processed through a GenAI-assisted internal peer review (Stage 6). The reviewer entered text-anchored marginal comments in the integrated draft (Task 6.1); a VBA macro converted those comments into inline, LLM-ingestible comments embedded in the body text (Task 6.2). The LLM then returned an iteratively revised draft with all changes shown in tracked changes (Task 6.3), and the reviewer accepted or rejected each proposed revision (Task 6.4). **QCG v** marked the point at which the model returned a revised draft requiring no further peer-review comments; where further comments remained, the process looped back to Task 6.1 for another cycle.

Final audit and finalization (Stage 7) used a purpose-built claim-and-citation audit pipeline (Task 7.1) within the "audit and verify" zone: the pipeline parsed every in-text citation, retrieved the corresponding extracted source passage, and used an LLM under a structured rubric to classify each cited claim as Supported, Partially Supported, or Not Supported. In the human audit and validation pass (Task 7.2), all Not Supported and Partially Supported claims were revised or removed, and a 10% random sample of the remaining Supported claims was manually re-verified against source PDFs. The audit was then re-run, with the cycle repeated until every cited claim returned a Supported verdict. A final editorial pass (Task 7.3) covered coherence and flow, formatting against the publication template, terminological consistency, and APA-compliance of the reference list.

A Brief (Recent) History of M&E System Conceptualization

Technical Conceptualizations

In the 1980s and 1990s, scholarship on monitoring and evaluation (M&E) tended to construe these activities in relatively narrow terms, centering on data management and accountability for development interventions. Such systems were largely treated as mechanisms to document inputs and outputs, ensuring compliance with donor requirements and the transparent use of public funds (Essis et al., 2023, p. 699).

For example, the World Bank's (1981) *Guidelines for the Design of Monitoring and Evaluation Systems for Agriculture and Rural Development Projects* defined monitoring as the provision of information to help management track implementation progress and make timely decisions, assessing whether inputs are delivered and used as intended and producing the expected initial effects. Evaluation, in turn, was defined as an assessment of a project's overall effects, both intended and unintended, and their impact (p. 4).

Project-specific M&E systems were conceived primarily as tools for reporting progress to project managers by providing regular, timely information to support implementation decisions and assess advancement toward project objectives (World Bank, 1981, p. 1). These interpretations cast M&E as a tool primarily for managerial oversight rather than a broader instrument for learning or strategic adaptation.

Casley and Kumar (1988, p. 2) added dimension to this technical vantage point by identifying three purposes of M&E data—description, explanation, and prediction—but noted that early practice typically emphasized descriptive data, answering "who, when, and where," without exploring causes. Data generated for physical and financial monitoring often served as primary examples of this descriptive emphasis, recording the relationship between expenditures and the achievement of targets without clarifying the reasons for progress or the lack thereof (Casley & Kumar, 1988, p. 2). While Casley and Kumar (1988, p. 3) advocated for wider applications of M&E data—such as examining beneficiary responses and assessing complex socioeconomic impacts—they acknowledged that these broader objectives were frequently constrained by their inherent technical complexity and the limited resources available for implementation.

This emphasis on accountability and project oversight persisted into the 1990s as formal M&E practices solidified within development agencies. Barton (1997, p. 10) characterized M&E systems as "a form of 'information system,' which is a broad term for information selection, gathering, analysis, and use. It can be described as a logical chain of linked ideas starting (and continuing) with information users." This linear model—data collection, analysis, and reporting—remained oriented toward donors and managers, treating evaluation as a periodic extension of routine monitoring. Historical analyses suggest that monitoring and evaluation, during this period, were largely "associated with accounting for development aid and answerability" (CLEAR-AA & TanEA, 2020, p. 4). Yet even as the technical focus prevailed, early references to "information users" suggested the latent potential of M&E data to serve broader objectives, notably by shaping knowledge-based decisions and fostering organizational learning.

Results-Based Management and National M&E Systems

By the 1990s and into the 2000s, the advent of Results-Based Management (RBM) expanded M&E's remit beyond basic compliance. Kusek and Rist (2004) reframed M&E as a results-based management tool designed to capture outcomes and feed evidence back into governance and decision-making processes. This shift moved the focus beyond traditional compliance—addressing the "did they do it?" question—to the critical "so what?" question, evaluating whether programs actually produced their intended results and how those findings could inform future policy corrections (Kusek & Rist, 2004, pp. xi, 12, 15). Practical guides, such as Kusek and Rist's (2004) *Ten Steps to a Results-Based M&E System* and Mackay's (2007) *How to Build M&E Systems to Support Better Government*, offered a blueprint for embedding M&E into government operations.

This period also saw a greater focus on the establishment of "national M&E systems," particularly in developing countries, as governments moved beyond project-level evaluations toward comprehensive, whole-of-government frameworks. Many countries established government-led institutional arrangements and central M&E structures designed to improve governance and foster greater accountability to their citizens, often formalized through specific evaluation policies reflecting a growing demand for well-structured frameworks supported by high-level political commitment (Essis et al., 2023, p. 700). In the mid-2000s, countries such as Ghana, Kenya, Rwanda, and Zambia emerged as early

adopters of structured national M&E systems, while nations like Benin, Uganda, Nigeria, and South Africa established national evaluation systems underpinned by dedicated policies and government-wide institutional arrangements (Essis et al., 2023, p. 700). These endeavors reflected a growing recognition of the importance of aligning M&E with a country's specific policy priorities, budget cycles, and accountability requirements. Mackay (2010) reinforces this point, contending that there is no universal model for an effective M&E system; rather, successful systems are contingent upon local factors such as government demand, the availability and quality of existing data, and the specific ways in which information will be integrated into the policy cycle (p. 4).

Consequently, M&E's purpose broadened to encompass organizational learning and performance management, providing data for strategic decisions, policy reform, and demonstrating results to citizens and donors. As this shift unfolded, the understanding of what constitutes a national M&E system began to mature, with successive definitions reflecting an increasingly holistic conceptualization. The Republic of South Africa (2008, as cited in Kanyamuna, 2019, p. 30) defined a Whole-of-Government Monitoring and Evaluation System (WoGM&ES) as a robust framework that provides an integrated and all-encompassing set of M&E practices, principles, and standards across government institutional structures, functioning as an apex-level system for information that draws from component systems to deliver essential M&E products tailored to the needs of users. That same year, Leeuw and Furu-bo (2008) characterized evaluation systems as institutionalized arrangements—defined by shared epistemological perspectives, organizational responsibility, permanence, and a focus on intended use—that provide a critical feedback mechanism regarding the outcomes and consequences of government action (pp. 159–160, 164). Engela and Ajam (2010) further elaborated this institutional dimension, characterizing a Government-Wide Monitoring and Evaluation System (GWM&ES) as a complex, multidisciplinary management framework designed to synchronize disparate departmental initiatives into a synergistic, composite system (pp. 28–29).

Alongside these institutional definitions, other scholars foregrounded the balance between the supply and demand of evaluative capacity as a defining feature of national M&E systems. Goldman et al. (2018, p. 1) defined a National Evaluation System (NES) as a "national system that guides how evaluations are selected, implemented, and used." The United Nations Evaluation Group (UNEG, 2022) emphasized the importance of balancing M&E supply and demand, proposing that "the supply and demand of evaluation capacity beyond government institutions (both in terms of evaluators and ECD offerings) is important for the national evaluation ecosystem" (p. 54). Fish et al. (2022)—drawing on Chirau et al.'s (2020) definition—characterize an NES as "the collection of M&E frameworks that guide the commissioning and undertaking of M&E, and the individual and institutional M&E capacity to use M&E information to inform development planning, policymaking, and decision-making" (Chirau et al., 2020, as cited in Fish et al., 2022).

Toward a Holistic M&E Ecosystem

Over the past decade and a half, the field has gravitated toward more inclusive, locally led, and learning-centric approaches, reframing M&E from a donor-oriented instrument into a comprehensive system for empowerment and change. Segone (2009, pp. 2, 9) characterized this shift as a "paradigmatic" change toward country-led evaluation (CLE), asserting that when national stakeholders own the entire process—from determining evaluation questions to applying findings—M&E functions as an "agent of change" rather than a tool for external accountability. Parallel calls to "decolonize" evaluation further underscored the importance of honoring indigenous epistemologies (Hopson et al., 2012, p. 62) and sharing methodological authority through equitable partnerships that center control over the initiation and interpretation of findings within the communities themselves (Johnston-Goodstar, 2012, pp. 111, 114).

This impetus for "country-led" systems prompted investments in institutional strengthening, recognizing that M&E is as much about governance structures and professional competencies as it is about data. Legislative frameworks and multisectoral partnerships expanded to anchor country-led M&E systems (GEI, 2023, pp. 3, 5), while the parallel pursuit of a "culture of evaluation" drew donors, civil society, academia, and local stakeholders into the joint production, dialogue, and use of evidence as a core component of good governance (Segone, 2009, pp. 23, 28).

As the field advanced, scholars and practitioners increasingly conceptualized M&E through the lens of "systems thinking," defining it as a complex whole of interacting and interdependent elements (Görgens & Kusek, 2009, p. 7). Under this framework, a functional M&E system operates as a management feedback loop where interlocking components—ranging from human capacity to data use—must all remain operational to ensure the system's overall effectiveness in informing governance and decision-making (Görgens & Kusek, 2009, pp. 2, 7). Central to this evolving conceptualization is the integration of feedback loops and adaptive management, driving the adoption of Monitoring, Evaluation, and Learning (MEL) frameworks. By positioning learning as a core objective, MEL frameworks transform monitoring and evaluation into a "strategic management and real-time learning tool," enhancing organizational capacity for continuous improvement and rapid responsiveness to shifting environments (Sartorius & Carver, 2006, p. 2). This marked a significant shift from traditional M&E focused on accountability to an approach that gives comparable, if not greater, weight to adaptive learning as a primary mechanism for maximizing program performance (Sartorius & Carver, 2006).

Beyond these structural and methodological considerations, more abstract factors—such as political will and evaluation culture—have also gained attention as constitutive dimensions of M&E systems. Lahey (2015, p. 2) argues that the development of national M&E systems depends on both political will and the technical capacity to build and sustain M&E infrastructure, while Lázaro (2015, p. 16) and Chirau et al. (2021, p. 1) emphasize that evaluation systems are embedded in the values, institutions, and governance contexts that shape how evaluative evidence is produced and used for policymaking.

In recent years, the term "ecosystem" has gained prominence as a way of capturing M&E's complexity and interconnectedness. Aggrawal et al. (2022, p. 23) define an evaluation ecosystem as "an overarching structure in which various types of organizations and partners in the field of evaluation interact, learn and grow together." They note that such ecosystems can function even without a national evaluation policy (NEP), as demonstrated by well-developed systems in Australia, the United Kingdom, and Singapore, though an NEP can enhance cohesion (Aggrawal et al., 2022, p. 24). Expanding on these definitions, Gates et al. (2022, pp. 1–2) advocate for an integrated, networked ecosystem model designed to catalyze systemic change through "new and emerging ways of evaluative practice." Overall, the growing ecosystem perspective reflects an expanded recognition that M&E rests on mutual influence and synergy among micro-level technical capacity, meso-level institutional frameworks, and macro-level sociopolitical and cultural contexts (Gates et al., 2022, p. 2).

Conceptual Frameworks for M&E Systems

Shifting conceptions of M&E—as evidenced by evolving definitions—have been mirrored in conceptual frameworks that aim to capture both the operational elements of M&E and the broader contexts within which they function. Early frameworks prioritized discrete technical components, whereas later models progressively integrated dimensions such as leadership, values, and stakeholder relationships, reflecting an expanding understanding of what constitutes an M&E system.

Holvoet and Renard (2007) offered one of the earliest comprehensive diagnostic frameworks, identifying six dimensions of M&E systems: policy, methodology, organization, capacity, participation of actors outside government, and quality and use of M&E (pp. 24–25). Their framework was predominantly technical, stressing clear policy directives, rigorous methodologies, and organizational structures to

support autonomy and impartiality, emphasizing that M&E services must remain "sufficiently autonomous from those that are responsible for implementation" to ensure reliability (Holvoet & Renard, 2007, p. 5). Similarly, Görgens and Kusek (2009, pp. 7–9) proposed a framework of twelve components organized into three concentric rings: people, partnerships, and planning; collecting, capturing, and verifying data; and using data for decision-making. Notably, by situating "advocacy, communication, and culture" within the outer ring of the "enabling environment," the authors explicitly moved beyond technical enumerations to acknowledge that a positive M&E culture—where evaluative activities are valued and encouraged—is essential for a system to function (Görgens & Kusek, 2009, pp. 227–228). Despite their publication over a decade and a half ago, these seminal frameworks remain relevant, continuing to influence contemporary scholarship and having been adapted or cited in works such as Goldman et al. (2018), Goldman et al. (2019), Ishmail and Tully (2020), and Chirau et al. (2021).

Building on these earlier foundations, Lopez-Acevedo et al. (2012, p. 22) identified three defining characteristics of successful systems: intensive utilization of M&E information, information quality and evaluation reliability, and system sustainability. While continuing to stress technical quality, they brought the concept of utilization further to the forefront, arguing that an M&E system's value lies in how its findings shape policies and decisions (Lopez-Acevedo et al., 2012, p. 5). Within this framework, sustainability is conceptualized as the system's capacity to endure beyond administrative transitions or changes in government leadership (Lopez-Acevedo et al., 2012, p. 7). Lahey (2015, p. 2) expanded upon such thinking by proposing an overarching framework consisting of two major influences—political will and technical factors—and four essential building blocks: leadership vision, enabling environment, capacity to supply and analyze information, and capacity to demand and use information. This integrated the core technical concerns of earlier frameworks with considerations of leadership and an "enabling environment" that ensures M&E findings can be incorporated into governance processes. Together, Lopez-Acevedo et al. and Lahey signaled a transition in how M&E systems are conceptualized: from cataloguing components to foregrounding the conditions—utilization, sustainability, political will—under which those components produce value.

Subsequent models ventured further by including intangible dimensions such as values, relationships, and ideology. Crawley (2017) introduced the "six-sphere framework," a diagnostic tool that conceptualizes M&E systems as a hierarchy of nested thematic domains moving from the explicit surface to the deep-rooted core: logistical, technical, contextual, relational, political, and ideological (pp. 1, 3). The logistical sphere addresses the availability of time and financial resources necessary to generate M&E data and facilitate its movement across the demand and supply chain (p. 3). The technical sphere encompasses the evaluation capacity of both producers and users, the quality of M&E products, and the fitness for purpose of the tools and systems used to channel and transmit information (pp. 3–4). The contextual sphere addresses the broader institutional environment, including organizational hierarchies, networks between stakeholders, and the "culture of evaluation" that determines how institutions learn from performance (p. 4). The relational sphere focuses on levels of trust and collaboration among stakeholders, recognizing that these interpersonal dynamics are essential for evaluation evidence to flow efficiently through the system rather than being diverted or discarded (p. 4). The political sphere examines whether high-level leadership and a vision for change exist to champion M&E, particularly in contexts where systematized evidence may challenge vested interests (p. 4). At the core, the ideological sphere delves into foundational values and principles—such as transparency, accountability, and equitable development—that must resonate with the state or institution for evaluation evidence to be meaningfully integrated into judgment (p. 5).

Though structured differently, Polastro and Prokop (2017, p. 31) converged on similar core elements. Their framework consists of three core dimensions—enabling environment, institutional capacity, and individual capacity—and seven key capacity factors: demand, use, leadership, resources, technical capacity, institutional arrangements, and values and standards. Their model is further reinforced by four cross-cutting principles—integration, inclusion, accountability, and evidence—that align na-

tional evaluation systems with the broader objectives of the 2030 Agenda (Polastro & Prokop, 2017, p. 31). Across these more recent frameworks, leadership emerges as a core enabling factor—evident in Lahey's emphasis on "vision of the leadership" and an "enabling environment" (Lahey, 2015, p. 2), and in Crawley's political and ideological spheres that foreground the values and principles conditioning whether evaluation evidence is welcomed and used (Crawley, 2017, pp. 4–5). This convergence reflects a broader move beyond purely technical components toward closer attention to the institutional and political context shaping M&E practice.

From a systems theory perspective, Stockmann et al. (2022) proposed a model that highlights three "subsystems": the political system (laws, regulations, parliamentary oversight), the social system (public discourse, civil society engagement), and the professional or academic system (education, professional associations, adherence to standards) (pp. 4, 19). Their approach reflects a structural focus, suggesting that an enduring M&E system must be woven into the fabric of governance, societal dialogue, and professional practice. Although they do not assign special categories to leadership or values, the inclusion of "public discourse" as a core component highlights the role of civil society as a "watchdog" and the necessity of community participation in legitimizing evaluative principles (Stockmann et al., 2022, pp. 4, 7). By arguing that state-led development alone is insufficient for the long-term anchoring of evaluation, Stockmann et al. (2022) reinforce the contemporary view that M&E institutionalization thrives only when supported by diverse sectors beyond the central government.

Most recently, Williams et al. (2024, p. 125) proposed a two-subsystem model that merges insights from preceding frameworks, arguing that evaluation capacity development (ECD), and by extension the broader M&E system, rests on capacity and capability subsystems. The capacity subsystem encompasses the "technical, organizational, political, and social skills and knowledge" required to design and undertake evaluations (Williams et al., 2024, p. 127). The capability subsystem focuses on developing the "technical, organizational, political, and social space" that allows those skills to be applied, essentially fostering the enabling environment necessary for M&E to flourish (Williams et al., 2024, pp. 127, 134). This distinction underscores that technical training (capacity) is only effective when the surrounding institutional culture and administrative structures provide the opportunity (capability) to use those skills meaningfully (Williams et al., 2024, p. 134). In this way, Williams et al. synthesize the trajectory traced across the frameworks reviewed here—from early technical cataloguing, through recognition of enabling conditions and leadership, to a unified model in which the capacity to do evaluation and the capability to apply it are treated as interdependent subsystems of a broader M&E architecture.

Conclusion

Taken together, these frameworks chart a discernible trajectory in M&E thinking. Earlier models, such as those by Holvoet and Renard (2007) or Görgens and Kusek (2009), identified robust measurement indicators and data systems as indispensable components for ensuring rigor and credibility. Görgens and Kusek (2009, p. 9) explicitly argued, however, that a system is only functional when this information is actively used to improve results—a final component they designated as the "bull's eye" of their framework. More recent formulations, including those of Lahey (2015), Crawley (2017), Polastro and Prokop (2017), Stockmann et al. (2022), and Williams et al. (2024), extend the scope to encompass political commitments, leadership vision, relational trust, cultural norms, and professional ethics. By highlighting how these intangible forces can either hinder or propel the institutionalization of M&E, they underscore the importance of viewing M&E as an ecosystem shaped by interdependent actors, overlapping systems, and shared – but evolving – values.

This trajectory invites reflection on prospective pathways. Crawley's (2017) framework – introduced about a decade after Holvoet and Renard (2007) – is itself approaching the ten-year mark, raising the question of whether future iterations of M&E system conceptualizations will diverge as significantly

from Crawley as Crawley did from Holvoet and Renard. Potentially, new frameworks could devote even greater attention to emerging issues, such as digital data ethics, artificial intelligence in evaluation processes, or real-time adaptive learning systems that capture complex and rapidly changing policy environments. On the other hand, it is also plausible that the field has reached a level of consensus on certain foundational principles – such as the interplay between technical capacity and political support – so that future frameworks refine or reconfigure these established elements rather than forging an entirely new paradigm.

The continuing use of earlier frameworks (e.g., Holvoet & Renard, 2007; Görgens & Kusek, 2009) in recent publications indicates their enduring value, even if in slightly updated forms. This sustained influence is neither inherently good nor bad; it can promote consistency and facilitate comparative analyses across contexts. At the same time, an overreliance on older models could, in theory, discourage innovation or blind practitioners to new considerations, especially as challenges like sustainability, equity, and rapidly advancing technology become more prominent. In this sense, the field of M&E faces an important balancing act: retaining the structured, time-tested insights of earlier frameworks while remaining open to evolving demands and cutting-edge methods. The question going forward may therefore be less about whether M&E systems need entirely new frameworks and more about how best to integrate emerging perspectives – be they technological, relational, or ideational – into robust, flexible models that can adapt to evolving contexts without losing the powerful insights garnered over decades of conceptual refinement.

References

- Aggrawal, R., Nandi, R., & Rao, B. V. (2022). Evaluation policy–The core of evaluation ecosystem. In *M&E@70: Strengthening India's evidence systems for accelerated reforms and inclusive growth* (p. 21). Development Monitoring and Evaluation Office. https://dmeo.gov.in/sites/default/files/2022-03/Compendium_of_Essays_new.pdf
- Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, 8(1), 19–32. <https://doi.org/10.1080/1364557032000119616>
- Barton, T. (1997). Guidelines to monitoring and evaluation: How are we doing? CARE – Uganda.
- Casley, D. J., & Kumar, K. (1988). *The collection, analysis, and use of monitoring and evaluation data*. The World Bank. <https://documents.worldbank.org/en/publication/documents-reports/documentdetail/534751468763750439>
- Chirau, T. J., Blaser-Mapitsa, C., & Amisi, M. (2021). Policies for evidence: A comparative analysis of Africa's national evaluation policy landscape. *Evidence & Policy*, 17(3), 535–548. <https://doi.org/10.1332/174426421X16104826256918>
- CLEAR-AA, & TanEA. (2020). Developing the Tanzania Evaluation Association visibility strategy: Scoping study report. CLEAR Anglophone Africa and Tanzania Evaluation Association.
- Crawley, K. D. (2017). The six-sphere framework: A practical tool for assessing monitoring and evaluation systems. *African Evaluation Journal*, 5(1). <https://doi.org/10.4102/aej.v5i1.193>
- Engela, R., & Ajam, T. (2010). *Implementing a government-wide monitoring and evaluation system in South Africa*. The World Bank. <https://documents.worldbank.org/en/publication/documents-reports/documentdetail/887221468115163336/implementing-a-government-wide-monitoring-and-evaluation-system-in-south-africa>
- Essis, E. M. L., Yaméogo, W. M. E., Sossa, O. G., Doukouré, D., Compaoré, R., Kpebo, D. O. D., ... Kouanda, S. (2023). Development of monitoring and evaluation systems in four national programs addressing mother and child health in Cote d'Ivoire: Qualitative analysis of the emergence and formulation process. *Risk Management and Healthcare Policy*, 16, 699–709. <https://doi.org/10.2147/RMHP.S377617>
- Fish, T., Dlakavu, A., Amisi, M., & Masvaure, S. (2022). Transforming voluntary organisations for professional evaluation into central pillars of national evaluation systems. *African Evaluation Journal*, 10(1), a608. <https://doi.org/10.4102/aej.v10i1.608>
- Gates, E. F., Madres, J., Hall, J. N., & Alvarez, K. B. (2022). It takes an ecosystem: Socioecological factors influencing equity-oriented evaluation in New England, US, 2021. *Evaluation and Program Planning*, 92, 102068. <https://doi.org/10.1016/j.evalprogplan.2022.102068>
- GEI. (2023). *GEI guidance on country partnerships*. Global Evaluation Initiative. <https://www.globalevaluationinitiative.org/>
- Goldman, I., Byamugisha, A., Gounou, A., Smith, L. R., Ntakumba, S., Lubanga, T., ... & Rot-Munstermann, K. (2018). The emergence of government evaluation systems in Africa: The case of Benin, Uganda and South Africa. *African Evaluation Journal*, 6(1), a253. <https://doi.org/10.4102/aej.v6i1.253>
- Goldman, I., Deliwe, C. N., Taylor, S., Ishmail, Z., Smith, L., Masangu, T., ... (2019). Evaluation² – Evaluating the national evaluation system in South Africa: What has been achieved in the first 5 years? *African Evaluation Journal*, 7(1), a400. <https://doi.org/10.4102/aej.v7i1.400>

- Görgens, M., & Kusek, J. Z. (2009). *Making monitoring and evaluation systems work: A capacity development toolkit*. The World Bank. <https://openknowledge.worldbank.org/handle/10986/2702>
- Holvoet, N., & Renard, R. (2007). *Monitoring and evaluation reform under changing aid modalities: Seeking the middle ground in aid-dependent low-income countries* (WIDER Research Paper No. 2007/52). United Nations University World Institute for Development Economics Research (UNU-WIDER). <https://www.econstor.eu/bitstream/10419/63293/1/546116027.pdf>
- Hopson, R. K., Kirkhart, K. E., & Bledsoe, K. L. (2012). Decolonizing evaluation in a developing world: Implications and cautions for equity-focused evaluation. In M. Segone (Ed.), *Evaluation for equitable development results* (pp. 59–82). UNICEF.
- Ishmail, Z., & Tully, V. L. (2020). An overview of the provincial evaluation system of the Western Cape Government of South Africa as a response to the evaluation of the National Evaluation System. *African Evaluation Journal*, 8(1), a425. <https://doi.org/10.4102/aej.v8i1.425>
- Johnston & Goodstar, K. (2012). Decolonizing evaluation: The necessity of evaluation advisory groups in Indigenous evaluation. *New Directions for Evaluation*, 2012(136), 109–117. <https://doi.org/10.1002/ev.20038>
- Kanyamuna, V. (2019). *Monitoring and evaluation systems: A meta-evaluation of the Zambian government's programme-based M&E system* [Doctoral thesis, Stellenbosch University].
- Kusek, J. Z., & Rist, R. C. (2004). *Ten steps to a results-based monitoring and evaluation system: A handbook for development practitioners*. The World Bank. <http://documents.worldbank.org/curated/en/638011468766181874>
- Lahey, R. (2015). *Devising an appropriate strategy for capacity building of a national monitoring and evaluation system*. World Bank.
- Lázaro, B. (2015). *Comparative study on the institutionalisation of evaluation in Europe and Latin America*. EUROSociAL Programme.
- Leeuw, F. L., & Furubo, J. E. (2008). Evaluation systems: What are they and why study them? *Evaluation*, 14(2), 157–169. <https://doi.org/10.1177/1356389007087537>
- Lopez-Acevedo, G., Krause, P., & Mackay, K. (Eds.). (2012). *Building better policies: The nuts and bolts of monitoring and evaluation systems*. World Bank.
- Mackay, K. (2007). *How to build M&E systems to support better government*. The World Bank. <https://openknowledge.worldbank.org/handle/10986/6851>
- Mackay, K. (2010). *Conceptual framework for monitoring and evaluation*. The World Bank. <https://openknowledge.worldbank.org/handle/10986/11080>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Peters, M. D. J., Godfrey, C., McInerney, P., Munn, Z., Tricco, A. C., & Khalil, H. (2020). Chapter 11: Scoping reviews. In E. Aromataris & Z. Munn (Eds.), *JBIManual for Evidence Synthesis*. JBI. <https://doi.org/10.46658/JBIMES-20-12>
- Polastro, R., & Prokop, M. (2017). Review of national evaluation systems and capacities for evaluating progress towards the Sustainable Development Goals: Emerging lessons learned from the Asia-Pacific region. In *People, planet and progress in the SDG era: Proceedings from the National Evaluation Capacities Conference 2017* (pp. 27–34). UNICEF & UNDP.

- Republic of South Africa. (2008). The role of premiers' offices in government-wide monitoring and evaluation: A good practice guide. The Presidency.
- Sartorius, R. H., & Carver, D. (2006). *Monitoring, evaluation, and learning for fragile states and peace-building programs*. Social Impact. <https://www.socialimpact.com/wp-content/uploads/2019/09/MEL-for-Fragile-States-and-Peacebuilding-Programs.pdf>
- Segone, M. (2009). Enhancing evidence-based policy-making through country-led monitoring and evaluation systems. In M. Segone (Ed.), *Country-led monitoring and evaluation systems: Better evidence for better policy and development results* (pp. 17–31). UNICEF.
- Stockmann, R., Meyer, W., & Szentmarjay, L. (Eds.). (2022). *The institutionalisation of evaluation in the Americas*. Springer Nature.
- Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., ... Straus, S. E. (2018). PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, 169(7), 467–473. <https://doi.org/10.7326/M18-0850>
- United Nations Evaluation Group. (2022). United Nations contributions to national evaluation capacity development and the evolution of national evaluation systems: An overview of implementation of General Assembly Resolution 69/237.
- Williams, B., Matt, E., Klier, S., Chaplowe, S., & Korswagen, S. I. (2024). Toward a systemic approach to evaluating evaluation capacity development. *New Directions for Evaluation*, 2024, 125–136. <https://doi.org/10.1002/ev.20612>
- World Bank. (1981). Guidelines for the design of monitoring and evaluation systems for agriculture and rural development projects. The World Bank.