global evaluation initiative

2.3

Measuring Evaluative Evidence Demand, Supply, and Use: A Scoping Review

Douglas Glandon
Guilherme Cunha Lima

Abstract

This scoping review examines how evaluative evidence demand, supply, and use are measured across diverse policy and organizational settings. It synthesizes academic and practitioner insights to clarify which indicators are most commonly used, the strengths and limitations of current measurement practices, and the contextual factors that shape how these dimensions are operationalized. This synthesis follows a scoping-review protocol grounded in JBI guidance, with reporting per PRISMA-ScR.

Measures of demand often include government or donor requests for evaluations, budget allocations for monitoring and evaluation (M&E), and formal policy frameworks mandating evidence-informed decision-making. Studies further distinguish among actual, potential, and latent demand, underscoring the role of institutional structures and donor influence, particularly in low-income settings.

Supply is gauged by assessing the size and composition of the evaluator pool, the availability of formal training pathways, and the strength of institutional networks and national evaluation systems. Quality indicators—such as evaluator competence, methodological rigor, and structured scoring frameworks—provide a further dimension, though concerns persist over the sustainability and depth of capacity-building initiatives.

Use is widely considered to be the most complex dimension to assess, encompassing instrumental, conceptual, and symbolic forms. Common indicators include the integration of evaluation findings into policy documents, stakeholder involvement in evaluation processes, and evidence of learning or behavioral change. At the system level, scholars have developed frameworks that assess whether institutionalized M&E architectures channel evidence into planning, budgeting, and accountability cycles. However, many metrics remain process-focused and do not capture long-term policy influence.

Cross-cutting themes include inconsistent measurement standards, donor-driven priorities, and the pivotal role of institutionalization in sustaining meaningful evaluation practices. The review concludes by calling for greater methodological standardization, more outcome-focused indicators, and longitudinal approaches that reflect the political and organizational realities influencing evidence use in policymaking.

Introduction

This scoping review addresses the research question: "What are the key measures used to characterize levels and changes in evaluative evidence demand, supply, and use?" The review synthesizes scholarship on how demand, supply, and use of evaluative evidence are conceptualized and measured across organizational, governmental, and sectoral contexts. The discussion proceeds through four major sections. First, it examines measures of demand, focusing on institutional markers such as budget allocations, national policies, and formal requests for evaluations. Second, it explores measures of supply, emphasizing indicators related to evaluator capacity, training programs, and institutional frameworks. Third, it addresses measures of use, describing how the integration of evaluation findings into policy and decision-making is tracked, including system-level frameworks that assess whether institutionalized M&E architectures are positioned to channel evidence into government processes. Fourth, it identifies cross-cutting insights and thematic observations related to data collection methods, standardization gaps, and the influence of institutional and political factors on all three domains. The review concludes by summarizing key findings and discussing implications for advancing measurement practices in the field of evaluation.

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 *metrics* sub-topic at the decision-maker/unit level.

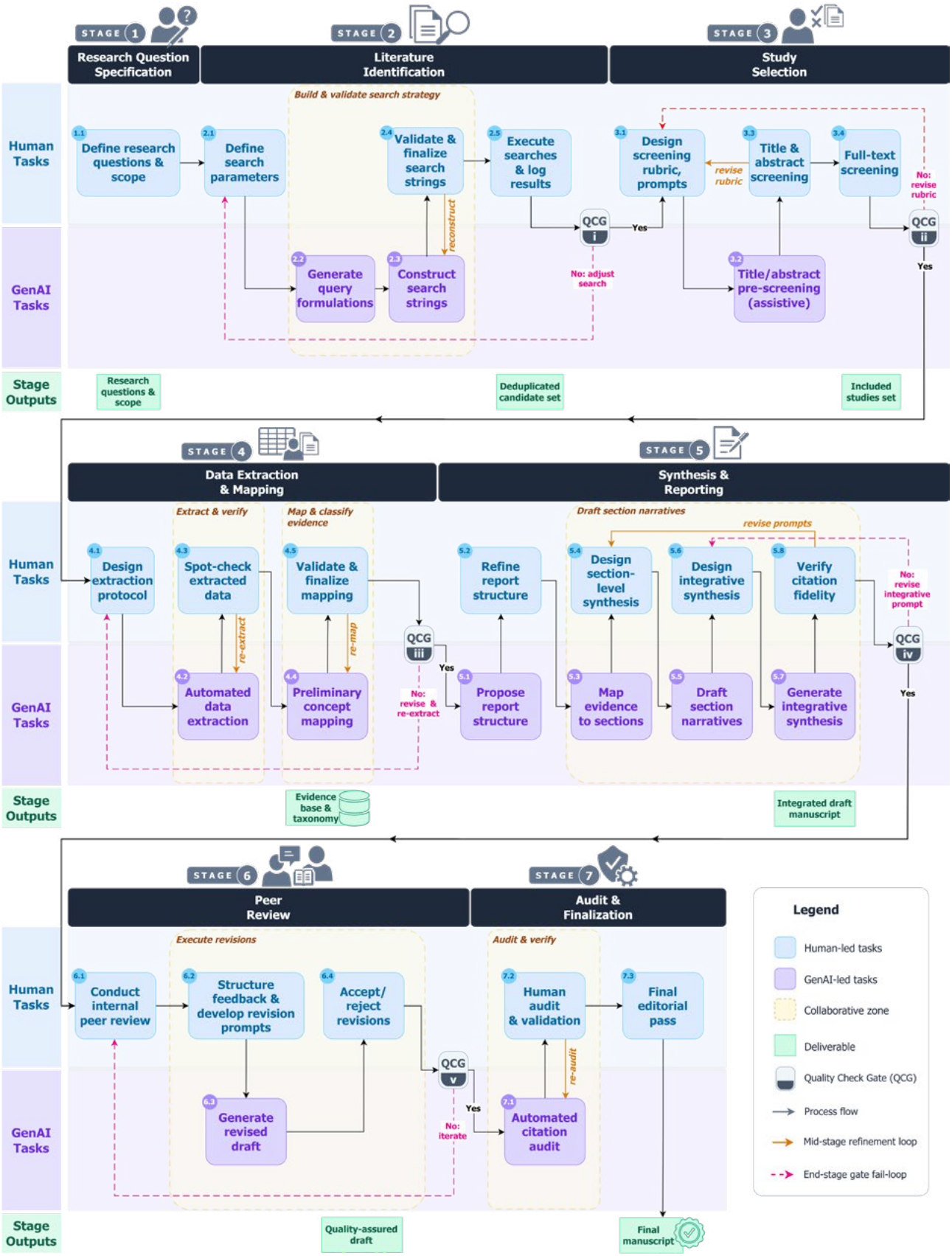
Methods

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 measurement-oriented and required breadth across heterogeneous, multidisciplinary material — peer-reviewed scholarship, instrument-validation studies, and institutional measurement frameworks and diagnostic tools. Eligibility followed the JBI Population–Concept–Context (PCC) framing, with the review question and inclusion criteria anchored primarily to the concept of measures and approaches for characterizing the demand for, supply of, and use of evaluative evidence. 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.

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



Eligibility criteria

Sources were eligible if they identified, defined, applied, or critically discussed measures, indicators, or assessment approaches used to characterize the demand for, supply of, or use of evaluative evidence — whether at the level of individuals, organizations, sectors, or countries. Eligible source types included peer-reviewed journal articles; instrument-validation studies; institutional measurement frameworks and diagnostic tools; and doctoral theses. No a priori date or language restriction was applied; included sources were entirely English-language. At full-text review, sources were excluded for not engaging substantively with measurement of demand, supply, or use, addressing measurement of unrelated phenomena, 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) describe, apply, or critically discuss a specific measure, indicator, or assessment approach for evaluative evidence demand, supply, or use, rather than invoking measurement 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).

Information sources and search strategy

Literature identification (Stage 2 of Figure 1) followed an iterative human–GenAI cycle, beginning with the human reviewer's definition of search parameters (Task 2.1). Scopus and Web of Science Core Collection provided peer-reviewed coverage, with Google Scholar searched as a complementary academic search engine. Four institutional repositories were systematically searched: the World Bank Open Knowledge Repository, the UNEG Publications Repository, the OECD iLibrary, and Global Evaluation Initiative (GEI) network publications. Because instrument-validation studies and assessment frameworks on evaluative-evidence use are frequently published in specialized institutional and professional-association channels, targeted searches of such channels — including the publication channels of regional evaluation partnerships and publisher channels hosting specialized measurement-focused evaluation-methods journal series — were conducted as a structured supplementary stream. Coverage was further extended through backward and forward citation chasing and expert recommendation.

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.

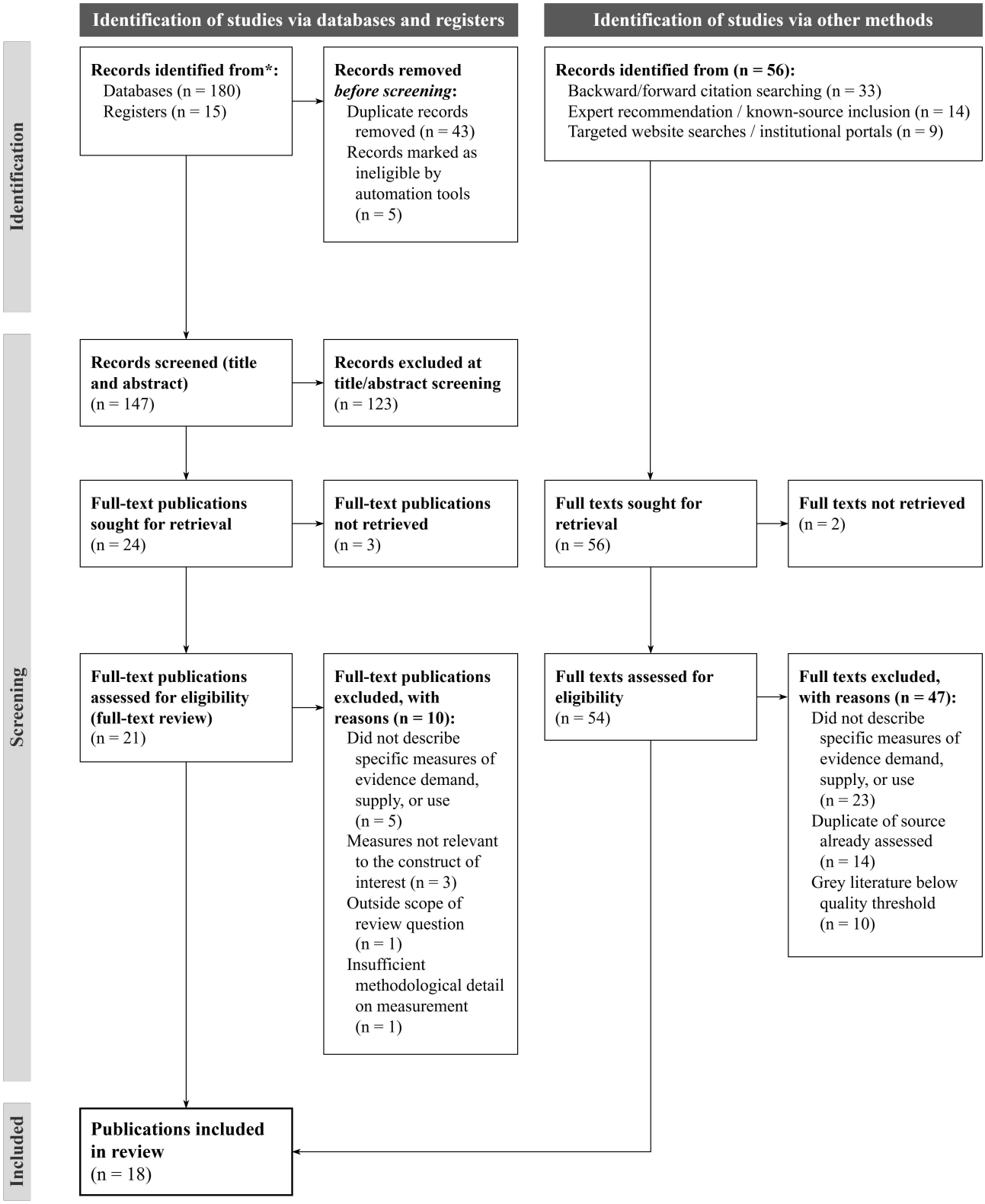
Source selection

Study selection (Stage 3) operated under a human-designed screening protocol (Task 3.1). A total of 195 records were identified through databases and registers (180 from databases; 15 from registers), with a further 56 identified through backward and forward citation searching, expert recommendation and known-source inclusion, and targeted website searches and institutional portals (**Figure 2**). After deduplication and metadata-based screening to remove records of document types outside the eligible scope, 147 records underwent title and abstract screening: the LLM produced assistive first-pass inclusion-likelihood labels (Task 3.2), and two independent human reviewers performed the title and abstract screening (Task 3.3), informed but not bound by the LLM labels. Twenty-four records advanced to full-text screening (Task 3.4); three could not be retrieved, leaving 21 assessed, of which

10 were excluded with coded reasons. In parallel, 54 of the 56 supplementary-stream records were retrieved and assessed (two were not retrievable), and 47 were excluded with coded reasons. A total of 18 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.

Figure 2. PRISMA Flow Diagram



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 measurement target (demand, supply, or use); the level of analysis; the specific indicators or instruments described; the type of measure (binary, quantitative, qualitative, composite); and the source's level of empirical grounding. 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 instrument-level extractions and measurement-target classifications 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), with mapping organized along the demand/supply/use scheme of the review and motivating three corresponding synthesis sections plus a cross-cutting section identifying shared themes and observations across the three.

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.

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 four-section analytical structure comprising three demand/supply/use synthesis sections and a cross-cutting section identifying shared themes and observations. 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.

Measures of Demand

Studies of evaluation systems across several African countries reveal varied measures for characterizing the sources and intensity of demand for evaluative evidence. A key conceptual framework in this literature is a tripartite classification distinguishing among "actual," "potential," and "latent" demand (Adams et al., 2013; Kumwenda & Latib, 2013). Adams et al. (2013, p. 7) define actual or effective demand as demand that is actually made, potential demand as demand that does not materialize due to lack of funding, and latent demand as demand of which there is no awareness but which represents a need for evidence that could be translated into demand. Both studies reference a shared Inception Report for detailed definitions and apply this framework to assess evaluation demand in their respective country contexts.

Applying this framework, studies from Ghana and Zambia find that actual government demand remains limited. Adams et al. (2013, pp. 27–28) identified limited actual demand from government in Ghana due to lack of awareness about evaluation's value, insufficient capacity to commission evaluations, and inadequate budget allocations—with monitoring activities accounting for approximately 63% of M&E expenditure while evaluation received far less. Similarly, in Zambia, Mulenga and Porter (2013, p. 32) found very little actual demand from government principals outside the development partner community, though potential entry points for demand exist within structures such as the Cabinet Office policy submission process. Many scholars note that donor-driven mandates often shape actual demand in lower-income countries. Mulenga and Porter (2013, p. 22) found that in Zambia most evaluations were commissioned and managed by development partners rather than being country-led processes, making external actors a strong driver of evaluation demand. Similarly, Kumwenda and Latib (2013) report that in Malawi, demand for evaluation is "largely conditioned by development partners," with development partners having "historically been the main source of demand for evaluations and for evidence based policy making" (pp. 11, 24). However, they note that the current government has taken positive steps to introduce a system of performance assessments and that there is growing evidence that the government is increasingly willing to support internally driven evidence-gathering exercises (Kumwenda & Latib, 2013, pp. 1, 24).

One frequently cited quantifiable indicator of demand is budget allocations for evaluative activities. Where national policies clearly mandate evaluations—by specifying, for instance, a fixed percentage of ministry budgets for evaluative activities—researchers treat these provisions as tangible measures of demand. In Ghana, Adams et al. (2013, p. 19) report that budget allocations for evaluation represented a small proportion of overall M&E expenditure, declining from 3% in 2009 to 1% in 2010 according to baseline data cited from NDPC (2011, as cited in Adams et al., 2013, p. 19). At a broader level, CLEAR-AA (2020, p. 15) identifies government initiatives to formalize M&E into government systems and anchor these systems in national budgets as key enabling factors for M&E capacity strengthening, suggesting that such institutionalization efforts support evaluation practice within government.

An additional set of measures involves specific institutional structures that signal demand. Adams et al. (2013, pp. 14–15) note that the Presidency in Ghana established a Policy Evaluation and Oversight

Unit (PEOU) in 2009 to monitor and evaluate government policies, programmes, and projects with the aim of generating evidence to guide decisions. CLEAR-AA (2020, p. 13) documents that countries such as Botswana, Tanzania, and Kenya are implementing institutional-level capacity strengthening by creating and capacitating M&E units within government ministries and departments as part of strategies to institutionalize M&E within government. A parallel indicator is the scrutiny by legislative or parliamentary committees, or mandates issued by central government agencies overseeing public expenditure. According to Niringiye (2018, p. 22), Uganda's National Policy on Public Sector Monitoring and Evaluation (2013) mandates that all projects valued at over 70 billion shillings must conduct rigorous evaluations—including baseline studies, mid-term reviews, and final evaluations—with project budgets required to allocate a percentage for evaluation as determined by the Development Committee. Niringiye (2018, p. 22) reports that approximately 28% of projects valued over this threshold represent the current government demand under this policy, illustrating how the percentage of projects requiring formal evaluation serves as a quantifiable measure of institutional demand.

Measures of Supply

Supply of evaluative evidence concerns the availability, capacity, and expertise of individuals and organizations capable of conducting rigorous evaluations. One foundational measure is the size and composition of the evaluator pool. As documented by Niringiye (2018, pp. 13–14), Uganda tracks the size of its evaluator pool through a roster developed by the USAID-funded Uganda Monitoring and Evaluation Management Services (MEMS) project, which in 2006 initially identified 75 individuals, 18 firms, and 11 institutions; this roster was updated in 2012 under the Mitchell Group's management to include 158 Ugandan experts in evaluation. This quantification offers a tangible measure of growth in capacity over time. In Zambia, Mulenga and Porter (2013, pp. ii, 2–3, 29) similarly identify a range of institutions contributing to the evaluation landscape, including an emerging array of think tanks conducting policy-relevant analysis, the Centre of Excellence at the University of Zambia, and consultancy companies and individual consultants with specialized expertise. Counting these institutions, assessing the qualifications of their staff, and examining their professional affiliations (e.g., membership in Voluntary Organizations for Professional Evaluation [VOPEs]) reveal trends in the national evaluation supply.

Beyond raw counts, researchers gauge the supply of evaluative evidence by examining the availability of formal training pathways—both university-level degree programs and intensive short courses. Wao et al. (2017, p. 2) note that nine universities in Kenya offer master's-level training in M&E, alongside intensive short courses aimed at upgrading professional skills. In evaluating one such short course, Wao et al. (2017, p. 3) developed a survey combining objective and open-ended items to assess participants' knowledge and skills in mixed-methods research, systematic review and meta-analysis, and core M&E principles, while also using a proxy indicator—participants' intention to change their professional practices post-training—to estimate potential improvements in job performance. More broadly, Masvaure and Fish (2022, p. 6) report that capacity-strengthening initiatives in several African countries commonly measure change through pre- and post-training scores, as well as by tracking the number of individuals trained and the number of graduates. At a more aggregate level, Donnelly and Searle (2017, pp. 310–311) discuss Labin et al.'s (2012, as cited in Donnelly & Searle, 2017, pp. 310–311) systematic review of evaluation capacity building (ECB) literature, which found that capacity-building outcomes are tracked at both individual levels (with 92% of articles reporting individual-level change) and organizational levels (with 77% of cases demonstrating organizational change), though Labin et al. acknowledged that specific outcome measures and quantitative data were inconsistently reported across the reviewed studies.

In addition to these input- and output-oriented training measures, quality indicators provide a further dimension of supply-side measurement. Johnson et al. (2009, p. 382) identify evaluator competence as

a characteristic under evaluation implementation, noting that six studies in their review addressed evaluator characteristics, suggesting that evaluation professionals play an important role in conducting evaluations that are used. Various frameworks assess the level of methodological rigor, the clarity of evaluation design, and timeliness of deliverables. Kanyamuna (2021, pp. 85–86) employs a LEADS scoring system in Zambia—adapted from Holvoet, Inberg, and Sekhritre (2013, as cited in Kanyamuna, 2021, p. 87)—which uses five categories ranging from "Little action" to "Sustainable," applied alongside a diagnostic checklist covering six dimensions: policy, methodology, organisation, capacity, participation of actors outside government, and use of M&E information. Although originally developed to evaluate both demand and use, LEADS can be adapted to measure the degree to which the supply side (e.g., public agencies, private firms) is institutionally ready to produce credible findings.

The distinction between local and international evaluators also serves as a critical supply-side measure, as multiple studies show that international consultants lead or dominate major evaluation exercises (Adams et al., 2013, pp. 4, 17; Kumwenda & Latib, 2013, p. 4; Niringiye, 2018, p. 7), a trend often attributed to perceived or actual gaps in local expertise (Adams et al., 2013, p. 10; Kumwenda & Latib, 2013, p. 4; Niringiye, 2018, pp. 5, 26). Hence, quantifying the ratio of local to international evaluators over time, as well as analyzing daily charge-out rates or labor markets for evaluators, can illuminate changing supply dynamics. At the institutional level, the presence of recognized M&E networks also contributes to evaluation capacity. Adams et al. (2013, p. 27) identify the Ghana Monitoring and Evaluation Forum (GMEF) as a voluntary organization for professional evaluators that acts as an interface between supply and demand for evaluations, advocating for evaluation use and providing training to enhance M&E practitioners' capacity.

Measures of Use

In the context of M&E systems strengthening initiatives, the notion of "use" encompasses not only whether the findings of an individual evaluation inform a discrete decision but also how an organization's evaluative architecture consistently channels evidence into successive planning, budgeting, and accountability cycles. Johnson et al. (2009, p. 378) report that the use of evaluation findings is traditionally divided into three types: instrumental, conceptual, and symbolic. Instrumental use captures situations in which evaluations directly inform decisions, for example, when a policy is revised based on new findings. Conceptual use occurs when stakeholders' understanding of a problem evolves due to exposure to evaluation results. Symbolic use occurs when a person uses the mere existence of the evaluation, rather than any aspect of its results, to persuade or to convince (Johnson et al., 2009, p. 378). Worton et al. (2017, p. 124) similarly describe how knowledge can be used persuasively to influence the thinking and decisions of others, such as funders or local decision-makers. Alongside these outcome-oriented categories, Johnson et al. (2009) identify stakeholder involvement as a distinct dimension, recognizing that including stakeholders in the evaluation process increases the likelihood of evaluation use (Donnelly & Searle, 2017, p. 309). This notion of process use—the learning or capacity gains stakeholders experience simply by engaging in an evaluation—may not be explicitly visible in final policy documents but can significantly shape organizational culture over time.

Measuring whether these forms of use actually occur requires tracking evaluation uptake across different institutional settings. Several studies document the extent to which evaluation recommendations are integrated into official policy documents or strategic plans. In Ghana, Adams et al. (2013, p. 27) find that actual demand for evaluation from the executive and legislature has been very limited. Most evaluations have been conducted by or on behalf of development partners, driven by donor requirements rather than government-initiated needs, and utilization of these evaluations by the executive, legislature, and government agencies has been limited (Adams et al., 2013, p. 17). Niringiye (2018, p. 10) describes how the Office of the Prime Minister (OPM) in Uganda works to promote the use of evidence in policy and programme formulation and decision-making through collaboration

with evaluation networks and expert institutions. However, Niringiye (2018, p. 25) also reports that staff involved in M&E at national and local levels indicate that systematic feedback loops of M&E outputs to sector and local-level planning and budgeting do not exist. These country cases illustrate a recurring gap between the formal aspiration to use evaluative evidence and its actual integration into government decision-making.

Beyond tracking the uptake of individual evaluations, some scholars have developed comprehensive monitoring and evaluation frameworks that specify indicators for measuring evidence use at scale. Kuchenmüller et al. (2022, pp. 2, 9–10) developed such a framework for the Evidence-informed Policy Network (EVIPNet) Europe that includes a theory of change and indicators across three key domains: knowledge translation (KT) capacity and skill building, network structure and governance, and KT/evidence-informed policy value and culture. The framework identifies 100 output and outcome indicators at the WHO Secretariat level and 52 at country team level, tracking measures such as KT products developed, capacity-building workshops attended by country team members and external stakeholders, and the extent of stakeholder engagement in knowledge translation activities (Kuchenmüller et al., 2022, p. 23).

At the system level, scholars have developed frameworks for assessing whether institutionalized monitoring and evaluation architectures are positioned to channel evidence into government decision processes. Gaarder and Briceño (2010, p. 301) use a comparative framework based on indicators such as reporting structure, organizational location, source of funding, and culture of utilization to assess the independence and policy influence of central evaluation bodies. Holvoet and Renard (2007, pp. 24–25) provide a diagnostic checklist that assesses whether M&E systems are structurally prepared to provide accountability and feedback, examining features such as the integration of M&E results into planning and budgeting, the incorporation of line ministries into centrally coordinated M&E functions (Holvoet & Renard, 2007, p. 24; Goldman et al., 2015, p. 4), the existence of formal governance mechanisms like parliamentary oversight (Holvoet & Renard, 2007, p. 25; Goldman et al., 2015, p. 5), and the establishment of feedback loops that link results to planning and budgeting cycles (Holvoet & Renard, 2007, p. 24; Goldman et al., 2015, p. 7). Goldman et al. (2015, p. 2), drawing on Holvoet and Renard's framework, provide one of the few systematic assessments of a national evaluation system, applying it to South Africa's National Evaluation System and tracking indicators such as the volume of government expenditure covered by evaluations (Goldman et al., 2015, p. 10), follow-up on recommendations through formal improvement plans (Goldman et al., 2015, p. 6), and the degree to which evaluation results inform national budget processes (Goldman et al., 2015, pp. 7, 10). Such system-level assessments move beyond measuring use of discrete studies toward evaluating whether institutionalized processes for iterative learning and evidence-based policymaking have been established (Goldman et al., 2015, pp. 2, 6).

Several authors offer structured checklists or rubrics to capture various dimensions of use through scoring or classification. Kanyamuna (2021, pp. 85–86), employing the LEADS scoring system, measures use of information from M&E in Zambia's government as a sixth diagnostic dimension, assigning scores from "Little action" to "Sustainable" based on whether M&E outputs are presented, analyzed, and employed as instruments of policy-making and advocacy. Similarly, Kanyamuna et al. (2018, p. 76) apply a four-component diagnostic checklist (originally developed by Bedi et al.) to assess the demand-side of Zambia's health sector M&E system, examining: analysis and evaluation; outputs and dissemination; integration with the budget process and parliament; and participation of non-state actors as demand-side stakeholders. Mansilla et al. (2024, pp. 1–2) develop a taxonomy of demand-driven questions by collecting and categorizing the types of questions that policymakers and other decision-makers commonly pose to evidence-support units, identifying 41 different types of demand-driven questions from more than 250 submitted examples. Taken together, these varying approaches underscore the complexity of measuring use: some methods rely on formal citation analyses, others adopt stakeholder surveys or interviews, and still others examine structural indicators, such as references to evaluation findings in newly enacted legislation.

Cross-Cutting Insights and Thematic Observations

A recurrent theme across the literature is the diversity of approaches used to measure evaluation demand, supply, and use. Studies employ varied approaches: Adams et al. (2013, p. 2) used a combination of desk review, including analysis of existing evaluation products, and semi-structured interviews with key informants, while CLEAR-AA (2020, p. 9) adopted a mixed-method approach comprising a cross-sectional survey and qualitative in-depth enquiry, noting that capacity-strengthening approaches by multilateral organizations vary from country to country (CLEAR-AA, 2020, p. 15). More structured measurement frameworks have also been proposed. As discussed in preceding sections, Kanyamuna (2021, p. 85) employed the LEADS scoring system—a five-point method with categories from "Little action" to "Sustainable"—adapted from Holvoet, Inberg, and Sekhritre (2013, as cited in Kanyamuna, 2021, p. 87), to assess Zambia's country-level M&E system. Worton et al. (2017, p. 125) propose the Community Knowledge Mobilization Evaluation (CKME) Framework as an approach to evaluating knowledge mobilisation designed to capture various forms of knowledge use within community settings. While LEADS provides a scoring system for assessing M&E system development at the country level, CKME focuses on how community stakeholders use knowledge from mobilization activities—together illustrating the diversity of structured approaches available. Despite these efforts, Donnelly and Searle (2017, p. 315), reviewing evaluation capacity building literature, observed that the ECB studies they reviewed provided very limited systematic reporting of the specific measures and quantitative indicators used, suggesting that even where measurement approaches exist, how they are documented and communicated remains inconsistent. CLEAR-AA (2020) reinforces this concern, noting a general dearth of literature on M&E system capacity development (p. 1) and documenting that survey respondents reported that process and output indicators were the only forms of measurement (p. 18), with measurement existing predominantly at the input, process, and output stages rather than capturing wider outcomes. More broadly, CLEAR-AA (2020, p. 17) notes that while some national systems attempt to coordinate monitoring through streamlined reporting frameworks, the actual measurement of capacity-strengthening effectiveness remains "amorphous and vague" (CLEAR-AA, 2020, p. 6) and is heavily influenced by the specific goals and accountability requirements of international funders (CLEAR-AA, 2020, pp. 6, 13). Johnson et al. (2009, p. 388) situate this challenge at a more general level, observing that systematic comparison across evaluation use studies is difficult because studies do not operationalize or measure the variables in the same manner.

A second cross-cutting challenge involves the political and institutional contexts that shape which aspects of demand, supply, and use are tracked. Donor-driven motivations for evaluation can skew measurement priorities, often emphasizing accountability over learning or long-term capacity. Kumwenda and Latib (2013) show that evaluation demand and use are frequently driven by external funders seeking accountability and resource monitoring (pp. 1, 9, 20), yet local stakeholders are unlikely to sustain these evaluative activities or fully internalize them into institutional decision-making if donor funding is withdrawn (pp. 4, 11). Similarly, Witter et al. (2017, p. 2) observe that evaluation evidence is often unavailable within the rapid decision-making windows required by political calendars, as in-depth evaluation processes may take several months or years to produce recommendations. This points to a tension between the supply side's emphasis on rigorous, time-intensive methodologies and the immediate operational evidence needs of policymakers (Witter et al., 2017, pp. 4, 9).

Underlying these challenges is a methodological divide in how data on evaluation demand, supply, and use are collected. Budget analyses, rosters of evaluators, and policy document reviews produce numerical indicators (e.g., percentage of projects evaluated, quantity of local experts, ratio of training programs), while interviews, case studies, and focus groups yield deeper insight into how and why evaluation findings drive policy (or not). The coexistence of these approaches underscores the difficulty of arriving at standardized metrics, as quantitative and qualitative methods capture different dimensions of the evaluation ecosystem.

Finally, the role of institutionalization appears repeatedly as a linchpin for sustaining meaningful measurement of demand, supply, and use. Studies indicate that institutionalized arrangements—such as national evaluation policies, mandated budget allocations for M&E, professional evaluation associations, and centralized mechanisms for consolidating and reporting evidence to decision-makers—are associated with stronger coordination of evaluation demand and supply and more systematic integration of findings into government processes (Adams et al., 2013, pp. 14, 28; Mulenga & Porter, 2013, pp. 10, 37). However, even in these countries, authors frequently note fragmentation between governmental and non-governmental M&E activities, underscoring the need for more unified frameworks to capture dynamic changes in the evaluation ecosystem.

Conclusion

In summary, the literature shows that measuring demand, supply, and use of evaluative evidence involves an array of indicators ranging from budget allocations and the presence of dedicated M&E policies or units (demand) to the capacity and expertise of evaluation professionals (supply) to whether findings meaningfully shape policies or practices (use). Many studies reference donor-driven measurements—such as the requirement for baseline, mid-term, and final evaluations—as key proxies for actual demand, but others emphasize more endogenous signals like parliamentary oversight and civil society advocacy. On the supply side, quantifiable metrics focus on the number and caliber of evaluators, the existence of national or regional M&E associations, and the prevalence of robust training initiatives. Use, meanwhile, is gauged not only through explicit policy uptake, citation in official documents, or evidence of process use via stakeholder learning, but also through system-level assessments that examine whether institutionalized architectures integrate evidence into planning, budgeting, and accountability cycles. Across the source material, the most robust measures appear to be those that combine structured frameworks (e.g., LEADS scoring, diagnostic checklists, M&E framework indicators) with both quantitative and qualitative insights on how evaluations are commissioned, conducted, and eventually utilized. Critically, where interventions seek to build durable M&E systems, future measurement frameworks should incorporate system-level indicators—such as the regularity of evidence reviews in budget negotiations or the proportion of cabinet dockets accompanied by evaluation syntheses—to complement study-specific uptake measures.

Looking ahead, these observations suggest a strong need for greater standardization in how demand, supply, and use are tracked, as well as more rigorous methodologies for linking evaluation outputs to longer-term policy outcomes. Notably, many of the documented measures remain process-oriented (e.g., counting evaluations, tracking training outputs, or noting the presence of M&E units) rather than capturing tangible shifts in decision-making culture or policy effectiveness. Future research could develop unified metrics that better account for the political and institutional contexts affecting commissioning behaviors, capacity development, and actual policy influence. Such efforts might include robust longitudinal studies that systematically compare evaluation systems within and across countries, while addressing political economy challenges that constrain the meaningful use of evaluative evidence.

References

Adams, S., Amoatey, C., Taabazuing, J., & Feinstein, O. (2013). *Study on the demand for and supply of evaluation in Ghana*. Centre for Learning on Evaluation and Results – Anglophone Africa (CLEAR-AA), University of the Witwatersrand.

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

Centers for Learning on Evaluation and Results – Anglophone Africa (CLEAR-AA). (2020). *M&E capacity-strengthening approaches and measurement in Africa* [Report]. https://unfccc.int/sites/default/files/resource/Clear%20ME%20Capacity%20report-V4b.pdf

Donnelly, C., & Searle, M. (2017). Optimizing use in the field of program evaluation by integrating learning from the knowledge field. *Canadian Journal of Program Evaluation*, *31*(3), 305–327.

Gaarder, M. M., & Briceño, B. (2010). Institutionalisation of government evaluation: Balancing trade-offs. *Journal of Development Effectiveness*, *2*(3), 289–309.

Goldman, I., Mathe, J. E., Jacob, C., Hercules, A., Amisi, M., Buthelezi, T., & Sadan, M. (2015). Developing South Africa's national evaluation policy and system: First lessons learned. *African Evaluation Journal*, *3*(1), a107. http://dx.doi.org/10.4102/aej.v3i1.107

Holvoet, N., & Renard, R. (2007). Monitoring and evaluation under the PRSP: Solid rock or quicksand? *Evaluation and Program Planning, 30*(1), 66–81. https://doi.org/10.1016/j.evalprogplan.2006.09.002

Johnson, K., Greenseid, L. O., Toal, S. A., King, J. A., Lawrenz, F., & Volkov, B. (2009). Research on evaluation use: A review of the empirical literature from 1986 to 2005. *American Journal of Evaluation, 30*(3), 377–410. https://doi.org/10.1177/1098214009341660

Kanyamuna, V. (2021). Towards building a functional whole-of-government monitoring and evaluation system for Zambia: The demand side. *World Journal of Social Sciences and Humanities*, *7*(3), 83–105.

Kanyamuna, V., Aurick, M., Emily, N., Chinyama, M., & Ackson, M. (2018). An assessment of the demand-side of the monitoring and evaluation system of the health sector in Zambia. *World Journal of Social Sciences and Humanities*, *4*(2), 75–86.

Kuchenmüller, T., Chapman, E., Takahashi, R., Lester, L., Reinap, M., Ellen, M., & Haby, M. M. (2022). A comprehensive monitoring and evaluation framework for evidence to policy networks. *Evaluation and Program Planning*, *91*, 102053. https://doi.org/10.1016/j.evalprogplan.2022.102053

Kumwenda, H., & Latib, S. (2013). *Study on the demand for and supply of evaluation in Malawi*. Centre for Learning on Evaluation and Results – Anglophone Africa (CLEAR-AA), University of the Witwatersrand.

Mansilla, C., Sweetman, A., Guyatt, G., & Lavis, J. N. (2024). A taxonomy of demand-driven questions for use by evidence producers, intermediaries, and decision-makers: Results from a cross-sectional survey. *Health Research Policy and Systems*, *22*, 78.

Masvaure, S., & Fish, T. E. (2022). Strengthening and measuring monitoring and evaluation capacity in selected African programmes. *African Evaluation Journal*, *10*(1), a635. https://doi.org/10.4102/aej.v10i1.635

Mulenga, O., & Porter, S. (2013). *Study on the demand for and supply of evaluation in Zambia*. Centre for Learning on Evaluation and Results – Anglophone Africa (CLEAR-AA), University of the Witwatersrand.

Niringiye, A. (2018). Diagnostic study on the supply and demand of evaluators in Uganda. Twende Mbele.

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.), *JBI Manual for Evidence Synthesis*. JBI. https://doi.org/10.46658/JBIMES-20-12

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

Wao, H., Onyango, R., Kisio, E., Njatha, M., & Onyango, N. O. (2017). Strengthening capacity for monitoring and evaluation through short course training in Kenya. *African Evaluation Journal, 5*(1), a192. https://doi.org/10.4102/aej.v5i1.192

Witter, S., Kardan, A., Scott, M., Lemma, S., & Samuels, F. (2017). Generating demand for and use of evaluation evidence in government health ministries: Lessons from a pilot programme in Uganda and Zambia. *Health Research Policy and Systems, 15*, Article 86. https://doi.org/10.1186/s12961-017-0250-4

Worton, S. K., Loomis, C., Pancer, S. M., Nelson, G., & Peters, R. D. (2017). Evidence to impact: A community knowledge mobilisation evaluation framework. *Gateways: International Journal of Community Research and Engagement, 10*, 121–142. https://doi.org/10.5130/ijcre.v10i1.5195