2026:1k

Sida Evaluation

Nordic Consulting Group Denmark

Final Report from the Evaluation of Sida's work with Poverty

Annex 4: How Can Evaluation Commissioners Think About Impact Assessments Going Forward?

Sida

Authors: Carsten Schwensen, Louise Scheibel Smed, John Rand, Anne-Lise Klausen and Ayla-Kristina Olesen Yurtaslan.

Sida Evaluation 2026:1k

Commissioned by Sida, Evaluation Team.

Published by: Sida

Date of final report: 2026-06-16

Art.no.: Sida62873en

urn:nbn:se:sida-62873en

This publication can be downloaded/ordered from www.Sida.se/publications

How Can Evaluation Commissioners Think About Impact Assessments Going Forward?

John Rand, University of Copenhagen – 26 May 2026.[1]

Introduction

Over the past five decades, rigorous evaluation has evolved from a niche, academically driven exercise into a central pillar of evidence-based policymaking. Early landmark studies in the 1970s demonstrated that experimental methods could generate credible causal evidence on the effects of public policies at scale. However, for much of the late 20th century, government performance was judged primarily on spending levels rather than outcomes (CGD, 2022).

The 1990s marked a turning point with the rise of the "results agenda". Governments increasingly demanded accountability for outcomes such as poverty reduction or employment, creating demand for stronger evaluative tools. This period laid the groundwork for what has since been termed the "evidence revolution" (CGD, 2022).

A major breakthrough came in the late 1990s and early 2000s with the application of randomized controlled trials (RCTs) to social policy. This demonstrated that rigorous causal evaluation was not only feasible in development contexts but could directly inform large-scale policy decisions. Over the following decade, a wave of specialized organizations and dedicated funding mechanisms emerged, institutionalizing RCT driven impact evaluation as a professional field (CGD, 2022).

Since the mid-2000s, the field expanded along three important dimensions. First, governments have increasingly embedded evaluation units within their own systems, reflecting growing political demand for evidence in decision-making. Second, the ecosystem has diversified to include "knowledge brokers" and evidence-to-policy intermediaries, aimed at translating research into actionable insights. Third, there has been a rapid growth in evidence synthesis products, such as systematic reviews and evidence gap maps, designed to make large bodies of research more accessible to policymakers (CGD, 2022).

More recently, the landscape has matured further with the rise of initiatives linking evidence to financing and global policy movements such as "What Works", all of which institutionalize the production and use of evidence in public systems (CGD, 2022). Taken together, this history reflects a clear trajectory: from measuring inputs, to demanding outcomes, to prioritizing causal attribution and policy relevance. However, it has also exposed a persistent tension. While many evaluations offer high internal validity, they are often costly, slow, and politically demanding. As a result, there is growing interest in complementary approaches that can produce credible, decision-relevant evidence without the full burden of long-winded evaluation designs.

Another concern is whether the rapid growth in evaluations meaningfully improved evidence generation and policy use in low- and middle-income countries? WIDER (2020) show that despite the rapid growth in evaluations, the use of evidence in policymaking is variable and often limited. Evidence is often not systematically tracked or integrated into decisions. The implication is that the increased production of evidence has not translated into proportional increases in use. However, impact evaluations do influence policy, but often indirectly. Evidence affects policy through multiple channels: (i) Instrumental use by direct program changes, (ii) Conceptual use by shaping broader policy thinking, (iii) Process use by building evaluation culture and capacity, and/or (iv) Symbolic use by supporting or legitimizing existing policies.

[1] Acknowledgement: The author is grateful to Ninja Ritter Klejnstrup (Chief Advisor; Evaluation, Learning and Quality (Learn); Danish Ministry of Foreign Affairs) and Carsten Schwensen (Senior Evaluator and Partner, Nordic Consulting Group Denmark) for valuable comments and suggestions that substantially improved this note.

An additional problem is that donor dominance continue to shape the evaluation landscape. Many evaluations are donor-funded and externally commissioned. This creates risks of weak local ownership, misalignment with national policy priorities, and limited domestic capacity development. Evidence suggest that sustainability and policy relevance improve when recipient countries themselves commission and finance evaluations (WIDER, 2020).

But maybe the most important insight from the past 20-30 years of evaluations is that synthesis products are critical for policy influence (WIDER, 2020). Systematic reviews and meta-analyses are more influential in shaping policy frameworks. Evidence gap maps (EGMs) help identify gaps and guide evaluation priorities as they present a visual overview of existing systematic reviews or impact evaluations in a sector or subsector, schematically representing the types of interventions evaluated and outcomes reported. Gap maps enable policy makers and practitioners to explore the findings and quality of the existing evidence and facilitate informed judgment and evidence-based decision making in international development policy and practice. The gap map also identifies key "gaps" where little or no evidence from impact evaluations and systematic reviews is available and where future research should be focused. Thus, gap maps can be a useful tool for developing a strategic approach to building the evidence base in a particular sector.

Implication is clear. The shift from single studies to aggregated evidence is essential for policy impact. But it is less clear at which stage in the process that synthesis products should be created; at the programme design phase or prior to a more "deep dive" evaluation? As stated in a CGD blog post from 2024[2] this "…*would include the use of all available evidence* ***at the design stage*** *… and subsequently identify where new evaluative evidence is needed. This would essentially entail starting with a comprehensive review of all the evidence on the instruments and key policy counterfactuals*…" during the programme design.

Across methodological guidance provided by international development organizations the central message is consistent: There is no single "best" approach to impact assessments. The appropriate design depends on the evaluation question, the intervention's causal architecture, and the constraints of time, data, budget, and institutions. For development programmes in particular, many interventions are not stand-alone causes but contributions that work only as part of a wider package of institutions, behaviours, policies, and contextual conditions. That is why mixed-methods designs and theory-led design choices matter so much (Vaessen et al, 2020).

An important operational distinction should nevertheless be acknowledged. Much of the impact-evaluation literature assumes relatively bounded interventions with identifiable beneficiaries, implementation timelines, and stable treatment definitions. Many commissioning agencies in development cooperation, however, increasingly evaluate broader strategies, thematic portfolios, financing modalities, country frameworks, or institutional partnerships rather than discrete projects. In such settings, evaluators often face diffuse intervention boundaries, overlapping causal pathways, multiple implementing actors, and evolving programme logics. This does not eliminate the relevance of impact-oriented thinking, but it changes what level of causal precision is realistically attainable and increases the importance of mixed-method, theory-based, and contribution-oriented approaches.

So literature suggests that the frontier is not a single method but a menu of options that become credible when three things are done well in advance: (i) the programme has an explicit theory of change, (ii) the available monitoring and administrative data are mapped before commissioning evaluations, and (iii) an evaluability assessment or equivalent design audit is conducted early enough to shape the study rather than merely justify it after the fact (Peersman et al, 2015).

A practical implication follows immediately. In most cases, the hardest problem is not lack of methodological knowledge. It is lack of design-readiness; no baseline, no treatment identifiers, no data-sharing agreements, no comparison logic, no agreed theory of change, and no prior synthesis of what is

[2] https://www.cgdev.org/blog/transforming-world-banks-approach-knowledge-rhetoric-reality

already known. When those foundations are missing, even sophisticated methods produce fragile findings. When those foundations are in place, several non-RCT options can produce evidence that is materially stronger than **stakeholder opinion** alone (Rogers, 2014).

A practical menu of options

Repeated quasi-experimental designs (difference-in-differences (DiD) type designs) are among the most useful "middle" options when the programme has a clear implementation date and repeated outcome data exist before and after implementation, ideally for both treated and untreated or later-treated groups. Their marginal cost can be relatively low when they rely on existing monitoring or administrative data, and they are often much simpler operationally than fielding a new survey-based RCT. Their main bottlenecks are weak or noisy routine data, unclear intervention timing, insufficient pre-intervention observations, and difficulty finding a comparison group unaffected by spillovers.

Threshold, rollout, and allocation-rule-based designs (regression discontinuity design (RDD), phased rollout comparisons, synthetic control methods) are often the strongest non-randomized options when programme rules or implementation realities already create a comparison logic. RDD is especially credible when eligibility is determined by a threshold and there are enough observations close to the cutoff. Synthetic control is particularly useful for area-level interventions when rich pre-intervention secondary data exist. Matching approaches can also be valuable, but only when one has rich covariate data on participants and nonparticipants; otherwise, unobserved selection remains a serious threat. These designs are simpler than an RCT in operational terms, but they are not "cheap tricks": they demand clean assignment rules, good documentation, a lot of data and competent analytical skills.

Repeated baseline-endline assessments combined with process evaluation are a useful alternative when a programme already has or can still collect pre-intervention and post-intervention data but cannot construct a convincing comparison group.[3] The literature is explicit that such designs should not be oversold as definitive causal attribution. Their value comes from improving on raw stakeholder perception by documenting change over time, implementation reliability, reach, exclusion, and identification of plausible mechanisms. They become much more credible when integrated with theory of change testing, triangulation, and monitoring-system data. In some portfolios, these designs are the most realistic option short of doing no impact-focused work at all (Dixon and Bamberger, 2022).

Theory-based and contribution-oriented approaches become relevant when interventions are complex, context-dependent, politically embedded, or difficult to evaluate using experimental or quasi-experimental designs. Rather than estimating a precise net effect, these approaches seek to develop causal claims about whether, how, and under what conditions an intervention contributed to observed change. Their strength comes from explicit theories of change, systematic testing of causal mechanisms, triangulation across data sources, and active consideration of alternative explanations.

This family of approaches includes contribution analysis, process tracing, comparative case studies, and participatory methods. Although these differ in emphasis and analytic strategy, they are often used in combination and are particularly useful when programmes are heterogeneous, sample sizes are small, outcomes emerge through multiple institutions or pathways, or context is central to explanation.

These approaches do not require a conventional control-group architecture, but they are not "light-touch" alternatives. Credible application requires careful case selection, transparent analytic procedures, systematic verification of claims, and teams capable of combining qualitative and quantitative reasoning.

And finally, *evidence synthesis products* (systematic reviews and evidence gap maps) are not substitutes for a programme-level impact evaluation, but they are a critical part of the options set because they determine whether a new evaluation is worth doing, where it will add the most value, and which design

[3] Process evaluation refers to assessing how and why a programme was implemented and functioned, alongside descriptive baseline-endline outcome assessment in the absence of a comparison group.

should be favoured. Systematic reviews tell commissioners what is already known and evidence gap maps show where strong evidence clusters already exist and where strategically important gaps remain. Campbell et al (2023) and Khalil et al (2025) shows, these products are increasingly central to research prioritization, funding choices, and evidence-informed decision-making.

Cost estimates and requirements

The most important point on costs is that a method's price is driven less by the label attached to it than by whether the design can piggyback on data and programme features that already exist. The cheapest credible options are usually those that exploit existing administrative and monitoring data, clear treatment flags, and observable rollouts. Costs rise sharply when evaluators need to create those foundations from scratch through surveys, retrospective reconstruction, or prolonged access negotiations (Gertler et al, 2016).

Administrative data are especially attractive because they can be less expensive and logistically easier than collecting new primary data, impose less burden on respondents, and often support long-run follow-up. But the hidden costs are significant. Data-use agreements for identifiable data typically involve substantial review; gaining access should begin in the design phase; providers may take months to extract and transfer files; and iteration over matching rules and file definitions is common. Even in relatively well-institutionalized settings, administrative files can carry nontrivial monetary fees (Rogger and Schuster, 2023), with J-PAL (2015) noting that administrative data extractions can easily be priced around US$10,000 per extraction.

Theory- and case-based approaches tend to shift cost from surveys and sample logistics toward senior analytical time, theory development, fieldwork quality, and synthesis discipline. The literature repeatedly describes these methods as specialist, time-consuming, and, if done well, resource intensive. That does not make them unattractive. It means they must be commissioned with the right skill mix and realistic timelines rather than treated as a cheap qualitative add-on. Their comparative advantage is not low price; it is that they remain feasible and policy-relevant where true counterfactual designs are blocked by ethics, politics, small samples, or complex causal chains.

Process evaluation is often the highest-return complement in this middle space. Dixon and Bamberger (2022) emphasize that process evaluation strengthens impact evaluation by revealing whether weak results reflect design failure, implementation failure, institutional constraints, or external shocks. It can also help explain effects, identify excluded groups, and feed back into programme improvement. In practice, the bottleneck is usually not conceptual acceptance but under-budgeting: commissioners want causal answers but do not allocate enough time or resources to examine mechanisms (Dixon and Bamberger, 2022).

Preparatory products are often underused because they are perceived as non-essential overhead, yet the literature suggests the opposite. Evaluability assessments usually represent a small share of total evaluation costs and can pay for themselves by preventing premature, underspecified, or data-impossible studies. Systematic reviews take time and evidence gap maps are also resource-intensive and vulnerable to coding subjectivity, but they are increasingly supported by automation and better software. In other words, these should not be optional in future impact assessments as they are the cost-effective architecture around later impact work (Khalil et al, 2025).

Binding constraints and possible ways forward

The first binding constraint is late commissioning. When the evaluation is designed only after implementation is underway, the programme usually lacks baseline measures, pre-randomization identifiers, treatment flags that can be linked to external data, and sometimes even a stable statement of who exactly the intervention was meant to reach. The consequence is predictable: weaker retrospective designs, greater ambiguity about whether groups were comparable, and limited ability to explain null results. The remedy is straightforward but organizationally demanding. Prospective design at programme inception, baseline or intake data collection, early treatment/comparison logic, and collection of identifiers before exposure so later linkage does not itself introduce bias (Gertler et al, 2016).

The second binding constraint is weak theory of change and indicator architecture. Many programmes have output indicators but lack intermediate, process, and mechanism indicators, or they use log frames that list activities and outputs without explaining how change is expected to happen. The result is an evaluation that can count activity but cannot diagnose where the causal chain broke. The literature is explicit that theory of change should guide data collection, analysis, and reporting, and that each element of intervention logic should be associated with indicators and data collection or verification methods. The practical fix is to map each causal step in advance and assign a preferred data source (M&E, administrative records, surveys, qualitative inquiry, and/or digital trace) to each step (Gertler et al, 2016).

The third binding constraint is fragmented or inaccessible data. Many donor-funded programmes could support stronger evaluation if their M&E records were linked to local administrative databases, but in practice those datasets sit in silos, use different identifiers, operate under unclear legal permissions, or are held by agencies that lack the staff time to support extraction and matching. The route around this bottleneck is not methodological cleverness alone. It is institutional work; data audits before commissioning, memoranda of understanding or data-use agreements, clear data-flow diagrams, agreed extraction schedules, realistic timelines, and willingness to support local data owners with staff time, in-kind analytical help, or other cooperation incentives. In some cases, on-site linkage at the data owner's premises is the most feasible compromise between access and confidentiality.

The fourth binding constraint is lack of a credible comparator. Programmes are often national, politically sensitive, or already rolled out everywhere. In those cases, the best response is not to default to generic qualitative opinion gathering. It is to look systematically for alternative comparison logic; rollout phasing, delays due to administrative factors, eligibility scores, geographic thresholds, synthetic controls, or natural experiments. If none exist, the evaluation should pivot honestly toward theory-based contribution analysis, process tracing, and comparative case work rather than pretending that net impact has been estimated. The constraint is not simply "no control group"; it is failure to adapt the research question to the level of causal leverage that is actually feasible.

The fifth binding constraint is capability. Administrative-data evaluation requires people who understand identifiers, linkage, data quality, privacy, extraction logic, and statistical assumptions. Theory-based work requires people who can specify a credible theory of change, formulate rival explanations, and synthesize mixed evidence. New digital methods require computational literacy and ethical judgment. The literature on government analytics and AI is clear that both data infrastructure and human capital matter. Without high-quality foundational data and the capacity to process, analyse, visualize, and interpret them, "analytics" becomes unreliable or performative. Capacity building is therefore not a separate reform; it should be part of the evaluation design itself (Rogger and Schuster, 2023).

A further constraint is organizational capacity within commissioning agencies themselves. Many evaluation units operate with limited staffing, broad thematic coverage obligations, and mandates that prioritize independence over embedded programme engagement. Under such conditions, extensive prospective evaluability work, detailed administrative-data mapping, or continuous technical engagement with implementers may simply exceed available operational capacity. This creates a

structural tension: the system increasingly demands more rigorous and data-intensive evaluations, while the institutional resources required to support such designs upstream remain limited. Future reform discussions should therefore distinguish clearly between what is methodologically desirable and what is administratively feasible within existing evaluation architectures.

What to expect of AI and big data for future impact assessments?

Bouyousfi and Ouedraogo (2024) convincingly argue that the most plausible near-term contribution of AI and big data is not to replace evaluation design but to expand the data and workflow options available to evaluators. They find growing interest in using big data and AI to capture and analyse social change and argue that the most promising path is to combine these tools with traditional approaches, leverage interconnected data platforms, mitigate ethical risks, and strengthen evaluator competencies in data and computer science. They also argue that interconnected data platforms can help address low response rates, missing data, and insufficient sample sizes in large-scale programmes, especially when paired with more conventional data sources. But the awareness about potential use cases should be acknowledged already at the programme design phase.

The strongest immediate use cases are text-heavy and synthesis-heavy tasks. IEG (2025) identifies at least five points in a structured literature review where large language models can help: document screening, extraction, annotation, summarization, and synthesis. But it is equally noted that generated outputs must be validated, manual review remains mandatory, and prompt workflows should be tested on human-labelled validation and test sets before broader use. In other words, AI is promising as an accelerator for evidence synthesis, transcript analysis, portfolio review, and coding of large text masses, but not as an autonomous source of trustworthy causal inference (IEG, 2025).

The risks are substantial and highly relevant for development evaluation. Bouyousfi and Ouedraogo (2024) highlight unresolved methodological, ethical, and ownership issues, including sampling and selection bias, uncertain representativeness, weak transparency, privacy concerns, and the danger that black-box models provide limited explanatory value for decision-makers. Their conclusion is notably balanced. Evaluators should use AI and big data with methodological flexibility, integrate them with traditional sources, improve ethics and governance, and invest in computational skills rather than assume that more data automatically means better knowledge. And the criteria for the use of AI should be outlined clearly by the agencies commissioning the evaluation.

That leads to a practical judgment for commissioners. AI and big data are best seen as an augmentation layer on top of sound evaluation design. They become genuinely useful only after the fundamentals are in place: an explicit theory of change, governance for data access and privacy, linked M&E and administrative data where possible, and human reviewers able to challenge the outputs. They do not remove the need for earlier data collection, stronger cooperation with local data owners, or better ToRs. They make those needs more important, not less (Rogger and Schuster, 2023 and Bouyousfi and Ouedraogo, 2024).

What commissioning agencies could change

The literature supports three conclusions on what commissioning agencies could change:

- **Programme designs should incorporate evaluation designs before implementation begins.** Prospective evaluations are stronger because they can collect baseline data, specify success measures during planning, and define a comparison logic before implementation changes the assignment process. The strongest practical improvement is to connect theory of change to a pre-implementation data map. For each causal step, define which indicators will be tracked, through which source, at what cadence, and whether those data can later be linked to local administrative records. Done well, this makes later quasi-experimental or theory-based designs far more credible (Gertler et at, 2016).
  In practice, however, this recommendation should be understood as an aspirational direction rather than a universally achievable standard. Many evaluations are commissioned years after programme initiation, often for accountability or strategic-learning purposes that were not anticipated at programme design stage. In such cases, the relevant question is not whether a fully specified evaluation design existed from the outset, but whether commissioners can retrospectively improve evaluability through clearer scoping, early reconstruction of programme logic, targeted data mapping, and more explicit prioritization of feasible causal questions. The practical challenge for many evaluation units is therefore less about implementing ideal prospective designs everywhere, and more about identifying where such investments are realistic and likely to generate substantial additional learning value.
- **Commissioning agencies should complete the data audit before writing the terms of reference (ToR).** The evaluation ToR is supposed to define objectives, scope, responsibilities, and the resources available for the study. The literature on evaluability and administrative data access strongly implies that data discovery should occur upstream, not be outsourced to consultants as an unfunded scoping exercise after contract award. A commissioning agency should therefore know, before tendering, what M&E data exist, what administrative data may be accessible, what identifiers and treatment flags are available, what legal permissions are needed, how long access is likely to take, and where the quality limitations sit. This will sharpen the ToR, improve bids, reduce wasted inception time, and make the eventual design choice much more defensible (IEG, 2011).
  At the same time, there are important practical limits to how much data discovery can realistically be completed before procurement. In many development programmes, access conditions, undocumented implementation changes, fragmented ownership of records, and evolving administrative systems mean that substantial elements of the evaluability assessment can only be completed once an evaluation team is in place and able to engage directly with implementers, data owners, and operational staff. In this sense, the inception phase serves an important operational function: not only refining methods, but also stress-testing whether the intended evaluation questions remain feasible under real-world data constraints. The key issue may therefore be less whether inception-phase data work occurs, and more whether ToRs specify sufficiently concrete expectations regarding what must be clarified, verified, and operationalized during inception.
  There are nevertheless important trade-offs associated with large stand-alone data audits. The expertise required to assess feasibility rigorously is often similar to the expertise required for the evaluation itself, creating potential procurement complications and risks of duplication. Moreover, teams conducting upstream scoping exercises may sometimes overestimate what can realistically be implemented once confronted with field realities, legal constraints, or data-quality problems during the evaluation itself. In some settings, a more realistic approach may therefore be hybrid models where commissioners undertake limited early-stage data mapping internally, while leaving deeper operational feasibility testing to the inception phase under clearer and more tightly specified contractual requirements.

- **Synthesis products and evidence gap maps should be commissioned before deciding on commissioning a new impact assessment.** Synthesis reviews and evidence gap maps answer broad questions, show where evidence clusters and gaps lie, and support research prioritization, funding decisions, and policy design. Recent work also shows that they can feed directly into more usable evidence products, toolkits and portals that summarize impact, cost, and evidence strength in a format decision-makers can actually use. That is a more effective route to future learning than commissioning isolated evaluations without first checking whether the evidence gap is real, important, and decision-relevant (Campbell et al, 2023).

Positioning four recent Sida Zambia evaluations against these recommendations

The four recent Sida Zambia evaluations (Diakonia, Musika, Beyond the Grid and Social Protection) confirm the core diagnosis above: the main problem is that evaluations are commissioned too late and on top of data systems that were not built for credible causal assessment. It is argued above that the real binding constraints are weak "design-readiness" and across the four studies, that is exactly what appears. All four reports rely on reconstructed theories of change, all four work heavily from secondary or pre-existing data, and all four are forced to make methodological compromises because key design choices were not locked in early enough. The result is not useless evidence; rather, it is evidence that is more fragile, narrower, and less diagnostic than it could have been.

The most important implication is that these reports should not be read mainly as a critique of evaluators. In all four cases, the evaluation teams seem to have done competent retrospective reconstruction under tight constraints. The deeper lesson is about commissioning practice: too often the system asks evaluators to recover impact answers at the end of a programme when the programme was never set up to generate those answers cleanly in the first place. This also implies a somewhat different understanding of evaluator competence than what is sometimes assumed in methodological guidance. In many real-world development settings, the evaluator's role is not merely to implement a pre-specified design on top of clean data infrastructure, but to identify credible learning opportunities under imperfect and evolving conditions. Strong evaluation teams are therefore often distinguished less by strict adherence to textbook designs than by their ability to combine methodological discipline with pragmatic adaptation, reconstruct feasible comparison logic, identify usable administrative and monitoring data, and generate transparent causal reasoning despite substantial constraints. But as stated above, prospective design, early data mapping, and an evaluability assessment are the highest-return reforms. The four evaluations provide concrete, applied proof of that claim.

The Musika study is the clearest example of an evaluation that found meaningful results, but on a data foundation that makes the causal story less secure than the headline findings might suggest. The report uses two main data sources: the Rural Agricultural Livelihoods Survey and Musika's Annual Household Surveys. Yet the two datasets do not align neatly with the core evaluation questions. In the AHS, the comparison group was "purposively" selected, and the documentation does not make clear how non-participant households were chosen. In the RALS, treatment households are not directly identified, the same households are not followed over time, and the study therefore has to rely on pseudo-panel methods. The report also states that it could not formally test the parallel-trends assumption for the reported difference-in-differences estimates because of limited pre-intervention data.

That matters because Musika is not a simple household-level treatment. It is a market facilitation programme working through firms, business models, and sector relationships. But the evaluation largely must infer those system-level effects from household and district proxies. This is why the study can identify some positive changes in income, food security, livestock, output-market access, and access to credit, while still being much less convincing on productivity, technology adoption, assets, spillovers, and systemic change. The report itself concludes that effects were largely concentrated around supported firms and districts, and that a more system-oriented market development approach was needed if Musika was to generate broader change.

As stated above, the binding constraint here was not lack of a method, but mismatch between the unit of intervention and the unit of measurement. For a programme like Musika, future evaluation should

not rely only on ex post household outcomes. It should begin with a partner-level registry, a clear geographic rollout map, consistent treatment identifiers, and a pre-specified plan for measuring both household outcomes and market-system change. Where rollout variation exists, a repeated quasi-experimental design is realistic. But because the core question is also about how markets change, that should be paired with process tracing or comparative case studies of firms, input markets, and spillover mechanisms. In other words, Musika is exactly the kind of case where the argument for mixed-method, theory-led design is most persuasive.

The Diakonia/Caritas SAP II evaluation illustrates a somewhat different but equally important lesson: even relatively ambitious mixed-method impact assessments remain constrained when the intervention logic, monitoring architecture, and treatment definitions were not designed prospectively for causal analysis. The evaluation team reconstructed the intervention theory of change ex post because no intervention-specific ToC had been developed for the Caritas component during implementation. The study therefore had to combine retrospective survey work, qualitative fieldwork, and quasi-experimental DiD estimates based partly on recall data, while openly acknowledging that key assumptions such as parallel trends could not be formally tested. The evaluation also documents fragmented baseline material, lack of outcome and impact indicators, incomplete beneficiary tracking, and limited disaggregation by gender and age. At the same time, the intervention itself evolved during implementation, with conservation farming and goat distribution added later as complementary livelihood components. This created overlapping and only partially documented treatment categories, making it harder to identify which programme elements actually drove observed changes.

Yet the evaluation is also a good example of how credible learning can still emerge under imperfect conditions when evaluators combine multiple methods transparently and remain explicit about limitations. The study identifies plausible positive effects on mining income, housing quality, protective behaviour, reduction of child labour, and some dimensions of women's decision-making, while simultaneously showing weak or contradictory evidence on conservation farming, goat distribution, diversification of livelihoods, and broader systemic change. Importantly, many of the weaker findings can be traced back to the exact design constraints discussed earlier in this note: late commissioning, insufficient monitoring systems, unclear treatment exposure, and lack of prospective data structures. The Diakonia case therefore reinforces the broader argument that the key bottleneck is not methodological knowledge alone, but whether programmes are made "evaluation ready" from the outset through explicit theories of change, stable treatment registries, baseline indicators linked to intended mechanisms, and data systems capable of tracking differentiated programme exposure over time.

The Beyond the Grid study shows an equally important failure mode: a programme can generate a large volume of operational data and still not be evaluation ready. The report notes that BGFZ placed substantial emphasis on data collection and analytics through the EDISON platform, with daily reporting from the energy service providers. Yet the evaluators could not access raw data because the platform was being phased out, transfer arrangements were ongoing, and the legal basis for access had not been established. As a result, the evaluation had to rely on already existing analyses, the earlier midterm evaluation, the ex-post evaluation, and stakeholder interviews. That is precisely the kind of fragmented and inaccessible data environment stated above as a core binding constraint.

The consequences are visible in the findings. The study is fairly strong on outreach and customer experience: it can credibly say that the programme reached more than one million people, that most customers were first-time users without a good alternative, and that perceived quality-of-life gains were substantial. But once the analysis moves up the theory of change, the evidence becomes noticeably thinner. The report explicitly says there is limited evidence on health and education impacts, no substantial assessment of women's employment and empowerment effects, and that long-term assumptions in the theory of change remain largely untested. It also finds that only a small minority of customers used energy productively, that most households remained on low-tier systems, and that last-mile reach remained difficult.

This is not mainly a story of programme failure. It is a story of evaluation design lagging behind programme design. BGFZ had a digital reporting backbone, but not a secured evaluation protocol

around that backbone. For future BGFZ-like programmes, the first reform should be contractual: raw-data access, data definitions, verification rules, and permissions for research use should be settled before implementation starts, not at endline. The second reform should be methodological: follow customer cohorts from onboarding, deliberately oversample poorer, more remote, and female customers, and pre-specify which higher-order outcomes the programme is realistically expected to influence. If the real strategic question is "does this model reach poorer and more remote households over time, and what kind of subsidy is needed to deepen reach?", then a repeated customer panel combined with process evaluation and comparison across rollout cohorts is a much better "middle option" than a loose ex post search for broad livelihood effects.

Among the four, the social protection study provides the strongest example of how much more useful an impact evaluation becomes when the design is even modestly better prepared. The study has a real baseline and endline for the pilot, a treatment and comparison group, multiple complementary datasets, and a transparent discussion of methods and limitations. Because of that, it generates not only positive findings, but genuinely diagnostic learning: better food security and dietary diversity, more gardening, some improvements in child health, stronger results in rural than urban areas, mixed labour effects, negative effects on prolonged breastfeeding and vitamin A uptake, and a significant negative pattern on asset accumulation. This is much closer to what can be described as credible decision-relevant evidence: not just whether something "worked," but which mechanisms worked, for whom, and where the programme logic broke down.

At the same time, this report also shows what remains missing when evaluation architecture is incomplete. The team states that it did not receive beneficiary payment histories from ZISPIS, could not access LCMS 2022 raw data, had to use pseudo-panels for RALS, and could not validate all critical difference-in-differences assumptions. It also finds that rural targeting was much stronger than urban targeting, which suggests that urban programming was scaled faster than the evaluation logic for urban contexts was developed. And at the national level, the report rightly notes that the transfer amount is far too small to expect large poverty-reduction effects on its own, which means that some evaluation questions were probably too ambitious relative to programme dosage.

Still, the contrast with the other three studies is instructive. Because this evaluation had better baseline, clearer treatment logic, and more mixed-method evidence, it could identify unintended harms and not just intended gains. That is a major lesson for future commissioning. A better impact evaluation is not one that simply produces stronger positive results. It is one that can credibly show where implementation, messaging, targeting, or programme design need correction. Future evaluations of national social transfer programmes should therefore build directly on administrative threshold rules, rollout timing, or payment variation; secure linkage between beneficiary records and payment, health, and nutrition systems; and treat urban targeting as a separate evaluability problem rather than assuming rural logic can be carried over unchanged.

Taken together, the four evaluations suggest five recurring mistakes in current practice. First, evaluations are too often commissioned after implementation has started, so the evaluator must reconstruct the theory of change instead of testing a prospectively specified one. Second, monitoring systems are treated as if they were evaluation systems, even though they often lack stable comparison groups, treatment identifiers, validated higher-order indicators, or permissions for linkage. Third, access to core data is not protected early enough through agreements with implementing partners and local data owners. Fourth, commissioners keep asking for hard impact statements even when the available comparator logic is weak, misaligned, or only partially credible. Fifth, process and mechanism analysis are underweighted, which means that mixed or null results are hard to interpret.

A particularly telling signal is that all four reports work from reconstructed theories of change. That should be read as an institutional warning sign. A reconstructed theory of change can be useful, but if evaluators repeatedly have to rebuild programme logic after the fact, then the commissioning system has not embedded evaluation where it belongs: at design stage. It is therefore right to put evaluability assessment, early data mapping, and evidence synthesis ahead of new impact studies. These are not side products. They are part of the minimum architecture for credible future evaluations.

The practical lesson is straightforward. For future Musika-like programmes, the best option is a mixed design built from the start: an explicit market-system theory of change, partner and district treatment registries, repeated outcome data for treated and later-treated areas, and a process-tracing component focused on market spillovers and firm behaviour. For future BGFZ-like programmes, the right design is a prospective customer panel built around platform data with contractual raw-data access, verified indicators on productive use and gendered outcomes, and comparison across rollout cohorts or underserved geographies. For future SCT-like programmes, the strongest "middle options" are threshold- or rollout-based quasi-experimental designs using PMT scores, enrolment timing, and payment histories, linked to health and nutrition systems and complemented by process evaluation on targeting and payment regularity. For Diakonia-like civil society and governance programmes, the implication is somewhat different because interventions are often locally adaptive, multi-component, and implemented through several partners and community structures simultaneously. In such settings, future evaluations should begin with a clearer intervention-specific theory of change, systematic beneficiary and activity tracking, and more consistent monitoring of treatment exposure across target groups.

At commissioner level, the recommendation can be stated even more simply. Before issuing a new impact-evaluation ToR, do five things: (i) commission a synthesis product/evidence gap map as early as possible in the process; (ii) complete an evaluability assessment; (iii) produce a data audit and secure data-sharing agreements; (iv) map each step of the theory of change to concrete indicators and data sources; and (v) decide honestly whether the feasible design supports net-impact claims, contribution claims, or both. The cost argument is clear: the cheapest credible studies are the ones that piggyback on data systems and rollout features that already exist, while retrospective reconstruction is what becomes costly, slow, and weak. Making this feasible requires more than early planning. It also requires investment in harmonized measurement and data systems. In practice, this means using stable identifiers, standardising core indicators and survey questions across waves and implementing partners, documenting metadata and coding structures, and designing administrative systems with evaluation use in mind from the outset. Small upstream investments in comparable measures and linkable data architectures often determine whether later evaluation can rely on existing systems or must resort to expensive retrospective reconstruction. The four evaluations give that principle real-world content. They show that what we are "doing wrong" is not mainly choosing the wrong estimator. It is waiting too long to make evaluation a design function rather than an endline reporting exercise.

Finally, not all development programmes warrant the same level of prospective impact-evaluation architecture. The likely learning value, scalability, strategic importance, innovation content, and expected future replication potential of an intervention should influence how much investment for an evaluation is justified upfront. Some programmes merit extensive baseline preparation and longitudinal tracking systems, while others may be more appropriately evaluated through lighter-touch contribution analysis, process evaluation, or synthesis-oriented approaches. The critical issue is therefore not methodological maximalism, but proportionality between evaluation ambition, operational reality, and decision-making needs.

References

- Bouyousfi, S.E. and Ouedraogo, M. (2024). Artificial intelligence and big data-driven evaluation research and practices: A systemic literature review. *Evaluation*, 31(3). https://doi.org/10.1177/13563890241289937
- Campbell, F., Tricco, A.C., Munn, Z. et al. (2023). Mapping reviews, scoping reviews, and evidence and gap maps (EGMs): the same but different - the "Big Picture" review family. *Syst Rev* **12**, 45. https://doi.org/10.1186/s13643-023-02178-5
- CGD (2022). A look Back at Two Decades of Progress in the Impact Evaluation Landscape. The Working Group on New Evidence Tools for Policy Impact. Center for Global Development.
- Dixon, V, Bamberger, M, 2021. Incorporating process evaluation into impact evaluation: what, why and how, Working Paper 50, New Delhi: International Initiative for Impact Evaluation (3ie). Available at: DOI http://doi.org/10.23846/WP0050
- Gertler, Paul J.; Martinez, Sebastian; Premand, Patrick; Rawlings, Laura B.; Vermeersch, Christel M. J. (2016). Impact Evaluation in Practice, Second Edition. World Bank.
- Goodrick, D. (2014). Comparative Case Studies, Methodological Briefs: Impact Evaluation 9, UNICEF Office of Research, Florence.
- IEG (2011). Writing Terms of Reference for an evaluation: A how-to guide. Independent Evaluation Group Report. The World Bank. Washington D.C.
- IEG (2025). Balancing Innovation and Rigor: Guidance for the Thoughtful Integration of Articifial Intelligence for Evaluation. Independent Evaluation Group (IEG) of the World Bank (WB) and the Independent Office of Evaluation (IOE) of the International Fund for Agricultural Development (IFAD).
- Khalil, H. et al. (2025). Methodology for mapping reviews, evidence maps and gap maps. Research Synthesis Methods, 16: 786-796. doi:10.1017/rsm.2025.25
- Peersman, G., Guijt, I., and Pasanen, T. (2015). Evaluability Assessment for Impact Evaluation. A Methods Lab publication. London: Overseas Development Institute.
- Rogers, P. (2014). Theory of Change, Methodological Briefs: Impact Evaluation 2, UNICEF Office of Research, Florence.
- Rogger, D. and Schuster, C. (2023). The Government Analytics Handbook. Leveraging data to strengthen public administration. IBRD/World Bank https://www.worldbank.org/en/publication/government-analytics
- Vaessen, Jos, Sebastian Lemire, and Barbara Befani (2020). Evaluation of International Development Interventions: An Overview of Approaches and Methods. Independent Evaluation Group. Washington, DC: World Bank.
- WIDER (2020). The impact of impact evaluation. WIDER Working Paper 2020/20.

The four evaluations used as case material:

- NCG/Sida (2025a). Impact study of Musika, a Zambian rural project. A case study as part of the Central Evaluation of Sida's work with Poverty.
- NCG/Sida (2025b). Impact study of Beyond the Grid for Zambia. A case study as part of the Central Evaluation of Sida's work with Poverty.
- NCG/Sida (2025c). Impact study of United Nations Joint Programme on Social Protection in Zambia A case study as part of the Central Evaluation of Sida's work with Poverty.
- NCG/Sida (2025d). The Strategic Evaluation of Sida's Work with Poverty. Case study of Diakonia's Strengthened Accountability Programme (SAP) II, Zambia.

Final Report from the Evaluation of Sida's work with Poverty

Main evaluation method: mixed-methods, synthesis analysis, literature review.

Positives: Across Zambia and South Sudan, Sida-supported interventions generated tangible benefits for poor populations, including improved incomes, food security, and access to essential services. In Zambia, interventions performed better as the systems are relatively stable while in South Sudan contributions have helped sustain basic services under extreme and fragile conditions.

Shortcomings: The impact results from Sida's work were uneven and remained limited in scale and transformative impact. Impact measures were often constrained by underuse of existing evidence in contribution designs and weak evaluability. Heavy reliance on partner-reported data, limited triangulation, and insufficient use of national data systems further created a verification gap and constrained deeper analysis.

SWEDISH INTERNATIONAL DEVELOPMENT COOPERATION AGENCY
Visiting address: Rissneleden 110, 174 57 Sundbyberg
Postal address: Box 2025, SE-174 02 Sundbyberg, Sweden
Telephone: +46 (0)8-698 50 00. Telefax: +46 (0)8-20 88 64
E-mail: sida@sida.se Web: sida.se/en

Sida