WFP EVALUATION

WFP
World Food Programme

SAVING LIVES
CHANGING LIVES

Annual Report from the Evaluation Methods Advisory Panel at WFP

2023 in Review

December 2023

Contents

Cover page photo credit: WFP/Sayed Asif Mahmud

Introduction

The Evaluation Methods Advisory Panel

Given the increase in the number of evaluations and the complex and diverse contexts in which the World Food Programme (WFP) operates, the WFP Office of Evaluation (OEV) has created an Evaluation Methods Advisory Panel (EMAP) to support improving evaluation methodology, approaches, and methods, and to reflect on international best practice and innovations in these areas. The Panel was launched in January 2022. Currently composed of six members (listed in annex 1), it complements provisions in the WFP evaluation quality assurance system (EQAS).

Purpose and Scope

The aims of the Annual Review are to:

- Reflect on evaluation approaches and methods used in evaluations, and progress towards improving and broadening the range of methodologies
- Identify systemic and structural challenges
- Derive lessons to increase quality and utility in future evaluations

The EMAP Annual Report covers most evaluations conducted by WFP's evaluation function – Policy Evaluations (PEs), Complex Emergency Evaluations (CEEs), Strategic Evaluations (SEs), Decentralized Evaluations (DEs), and Country Strategic Plan Evaluations (CSPEs) – in 2022-2023 (see Annex 3). It is based on reviews undertaken by EMAP members ("the reviewers"), and discussions and workshops between the reviewers and WFP. EMAP has not examined system-wide and impact evaluations.

Process

Two approaches to the EMAP reviews were undertaken. In one strand of activities, EMAP members received a selection of completed CSPE and DE evaluation reports (ERs), and the related terms of reference (ToR) and inception reports (IRs), for their review. The other strand of EMAP activities was giving feedback on draft outputs for Policy Evaluations (PEs), Complex Emergency Evaluations (CEEs) and Strategic Evaluations (SEs).

Two EMAP advisers wrote this Annual Report; the process of preparing it entailed:

- Review of the advice provided by EMAP on WFP evaluations during 2023.
- Discussion of the draft annual report with OEV, Regional Evaluation Officers (REOs) and other EMAP advisors in a two-day workshop at WFP. This report incorporates key elements from these discussions.

As in 2022, the 2023 review faced the following limitations:

- The review included 14 DEs, 10 CSPEs, 5 PEs, 3 CEEs and 3 SEs, but analysed outputs were at different stages of development. EMAP reviewers prepared review reports for DEs and CSPEs based on *finalised* ToRs, inception and evaluation reports. Conversely, for SEs, PEs and CEEs, the reviews examined *draft* concept notes, ToRs, IRs, ERs, and two literature reviews.
- Not all EMAP reviews undertaken in 2023 were finalised in time for the synthesis process undertaken to prepare the Annual Report.

- Most reviews followed a structure provided by WFP which varied by evaluation type. For instance, the DE review template included a section on overall evaluation approaches and methods which was not included in the CSPE review template. Some reviews did not use the templates provided but added comments directly to the draft reports.
- Finally, reviewing written evaluation outputs presented challenges to explaining why something did or did not happen in an evaluation process.

Unlike in 2022, there was no opportunity for the EMAP to discuss the draft annual report as a panel before sharing it with OEV. The 2023 Annual Report was, however, discussed in a workshop with EMAP members and OEV staff, including regional evaluation officers, to validate the results and discuss potential ways forward across the different types of evaluations in WFP.

Structure of the EMAP Annual Report

The following sections will explore issues within themes selected by OEV based on the initial feedback from the EMAP review of evaluation deliverables:

1. Evaluation approaches and methods
2. WFP guidance for robust and creative evaluation design
3. Use of theory-based evaluation
4. Linkages between elements of the evaluation design
5. Triangulation, clarity, and transparency
6. Lessons to strengthen WFP's evaluation function

Each section starts with the question set by OEV for the report, examines the current practice and challenges in response to the question, and identifies good practice that can be leveraged as well as avenues for further exploration by OEV.

WFP/Hebatallah Munassar

1. Approaches and methods

To what extent are the approaches and methods applied across the evaluations similar or do they include innovative elements? What are some good practices which could be leveraged to enhance the design of evaluations?

Overview of current practice across evaluations

Compared with the evaluations EMAP reviewed in 2022, the 2023 evaluations display a broader methodological range. For instance, **evaluation approaches** in DEs have included among others, utilization focused evaluation, developmental evaluation, quasi-experimental design with propensity score matching, and elements of equity-focused and participatory approaches. Most CSPEs claimed to use a theory-based approach to the evaluations (see section 3).

Analytical methods have included aspects of Contribution Analysis, Outcome Harvesting, Process Evaluation or Process Tracing, Qualitative Impact Protocol, Social Network Analysis, Most Significant Change and tools borrowed from other sectors, such as the Kirkpatrick model to assess training, and a framework to assess change in market systems. A few evaluations used rubrics to define criteria for performance levels and to visualize findings in the form of heat maps. However, in many evaluations, methodology sections were largely focused on data collection tools and gave insufficient attention to analysis frameworks and methods.

Other methods were recommended in the ToR, but not applied by evaluation teams without any explanation as to why (e.g., Qualitative Comparative Analysis, QCA). In some cases, the IR referred to a specific method, but the evaluation report lacked evidence for the use of that method (e.g., contribution analysis). Despite WFP's efforts to mainstream gender, only few decentralized evaluation reports included forms of **gender analysis.**

Data collection largely relied on methods of social research, usually document review, key informant interviews (KIIs), (focus) group discussions (FGDs), standardized surveys (online, phone, and in-presence), and some direct observation. One team used real-time polling in online FGDs. Some CSPE made attempts to increase efficiency in data collection by adding specific questions to existing surveys. Case studies were also used in some CSPEs and DEs.

Commonalities and divergences

Common features across the evaluations were mixed methods approaches and classical social research tools for data collection. Most CSPE and DE reports included the (re-) construction or visualisation of a theory of change (ToC), although TOC use in the evaluation varied (see section 3). Even when ToRs stated that learning was a main purpose of the evaluation, the vast majority of evaluation questions was mainly aimed at accountability, asking about the extent to which a certain criterion or target was met. Correspondingly, evaluation findings and conclusions usually emphasised the evaluand's performance against plan, offering limited insights for learning from gaps, mistakes, and failure. ERs tended to pay little attention to unintended results, especially where the ToR did not include any question on that subject.

Finally, evaluations continued to respond to contact restrictions linked to the Covid-19 pandemic, commonly with adjustments in evaluation teams, itineraries, and data collection methods (e.g., online and hybrid formats, smaller in-presence FGDs).

Divergences were found regarding the clarity and transparency of evaluation design as described in IRs and ERs. The user-friendliness of evaluation reports also varied. Some ERs presented their findings clearly; in others, findings appeared to be hidden behind raw data (quotes, statistics), making it hard to discern the evaluators' analysis.

Systemic challenges

How can WFP foster methodological innovation while ensuring all necessary quality standards are met? Close WFP guidance, as exemplified in the Decentralised Evaluations Quality Assurance System (DEQAS), ensures all evaluations reach a certain level of quality, but can stifle creativity. Limited time for field research (e.g., 2-3 weeks in DEs) can persuade evaluators to resort to familiar methods, since field testing and piloting new tools might be impossible. CSPEs also have a clearly defined scope and face some restrictions on what can be achieved with in the timeframe and budget. Conversely, Strategic Evaluations, covering a wide range of different themes, are more open and with sufficient interest and time, evaluation teams can be more innovative.

Across evaluations, the evaluation questions (EQs) usually cover all six OECD-DAC criteria, even though OECD-DAC advises selecting a purposeful set of criteria from these.[i] Where evaluation questions are too broad or too many, breadth of research is privileged over depth. Also, evaluations that address dozens of EQs and sub-questions tend to generate long lists of recommendations that might overwhelm project and programme teams.

Good practices to be leveraged / explored

Several evaluations displayed examples of good practice that could be leveraged when discussing evaluation approaches and design with evaluation teams.

- Tailoring a manageable set of evaluation questions to (i) the purpose of the evaluation (institutional learning vs. performance assessment) and to (ii) the evaluand – a practice encouraging evaluation teams to develop the appropriate evaluation design. Where theory-based approaches are used, ensuring that assumptions are reflected in the set of evaluation questions
- Developing conceptual frameworks which will structure the evaluation analysis.
- Literature review preceding the development of the inception report, grounding the evaluation design in a strong understanding of its context
- Gender and equity analysis for a nuanced understanding of gender and other equity-related dynamics

Two DE teams decided against using the method (QCA) proposed in the ToR, which was good practice in those cases, as the resource framework for the evaluation would have ruled out proper application of that method. As a rule, evaluators should be encouraged to critically engage with ToR and propose only methods that fit into the resource framework.

2. Evaluation guidance

To what extent should WFP evaluation guidance be strengthened and enhanced to help improve evaluation design, while encouraging a flexible and adaptive approach to increase innovation/ creativity, rather than compliance?

Overview of key areas for which guidance can be strengthened & standardized

The WFP Office of Evaluation provides guidance to evaluation managers and evaluation teams. The level of detail varies with the type of evaluation. DEQAS provides most detail, as country offices do not necessarily have evaluation specialists. Detailed guidance ensures that all evaluations (centralized and decentralized) use ToR, IR and ER templates that include all important aspects (compliance). But the quality of the evaluations following those templates still varies greatly, and innovative evaluation designs are rare.

Potentially, WFP can foster flexible, adaptive, and innovative approaches by combining a necessary (but limited) degree of regulation and standardisation with purposeful dialogue. For example, early in the inception phase, evaluation managers (and possibly programme officers in DEs and CSPEs) should discuss all evaluation questions with the evaluation team to clarify what they mean in the specific evaluation and define key terms together. Alternatively, standard definitions of key terms, such as equity or HDP nexus, should be included in evaluation ToRs, also to help broader learning and synthesis efforts. Based on shared, documented definitions, and an encouragement to be creative, evaluation teams can develop an appropriate methodology. Mixed-methods and theory-based approaches are often useful, but not always needed.[ii]

The informal practice of sharing a good evaluation report of similar intervention as a model for the evaluation team should be discontinued, as it may keep evaluation teams from developing the design that works best for the specific evaluand and the specific EQs.

Systemic challenges which limit adaptability / innovation / flexibility

A common issue is the short timeline between evaluation commission and delivery of the evaluation report, especially for some decentralized evaluations, which implicitly discourages evaluation teams from tailoring the evaluation to the specific needs and from trying out creative approaches.

Despite their importance for OEV, there are a number of thematic areas where the discussion in many evaluations is a little thin and where conceptual frameworks are required to organize inquiries. For example, CSPEs found this to be the case when examining issues related to equity (beyond gender) and inclusion. Simple guidance, including definitions and potential lines of inquiry for these issues, needs to be prepared.

Interaction between evaluation managers and evaluation teams tends to be limited to the production of key deliverables, often focusing on feedback loops on IR and ER drafts. More continuous interaction, with joint reflection before the IR is drafted, and light touch communication during the data collection and analysis phase, is recommended. That could ensure that WFP and evaluation teams can base their work on a shared understanding of the evaluation purpose, its questions, challenges and ways in which the ET has dealt with them.

Good practices to be leveraged / explored

Evaluation planning, including ToR development, and the inception phase are the moments where WFP is in the best position to influence evaluation quality. DE Terms of Reference reviewed in 2023 tend to be clearer and user-friendlier than those reviewed in 2022, reflecting improved WFP guidance.

3. Use of theory-based evaluation

How could the theory-based evaluation approach & theories of change be more meaningfully designed, implemented and used to strengthen the evaluation design?

Overview of current practice across evaluations

Most of the 10 CSPEs examined stated that they took a Theory Based Approach (TBA) but not all clearly defined what it meant or how it was applied in practice. Two evaluations did not mention use of a TBA, but they did take steps consistent with such approach, including the development of a Theory of Change. In most sampled CSPEs there was no pre-existing ToC, so one was constructed by the evaluation team during the Inception Phase[iii].. All reviewed DEs referred to an existing or reconstructed theory of change, usually briefly described, and sometimes assorted with a list of assumptions. A few DEs displayed elements of a theory-based approach, others used their ToC only for illustration, if at all. TBA is also used in some PEs; for example in identifying possible causes enabling and limiting the implementation of the policy.

Key gaps and systemic challenges

- DEs and CSPEs often referred to or reconstructed theories of change, but rarely used them for causal analysis. Most inception and evaluation reports did not describe the rationale for and steps of using a theory of change. There is a slight disconnect between the expectations of WFP's standard template of evaluation questions and the normal thrust of theory-based evaluation enquiries. For example, only one of the evaluation questions in the CSPE template specifically deals with the contents of the Theory of Change (2.1 i.e. *To what extent did WFP deliver expected outputs and contribute to the expected country strategic plan strategic outcomes? (Effectiveness)*). The focus of the two subsections is on documenting the achievements at the output and outcome level and not on causal analysis, which requires a more vertical perspective looking at the sequence of events connecting inputs activities outputs and outcomes and impact. Though some attention is directed under "lines of Inquiry" to "*Describing logical connection between activities implemented and outputs*" and the same at the outcome level. EQ4 from the same template looks like it might address this problem, when it asks "EQ4: *What are the factors that explain WFP performance and the extent to which it has made the strategic shift expected by the country's strategic plan?*" But in practice all the sub questions are about high-level generic questions, about use of evidence, ability to mobilise resources, partnerships, and flexibility in the face of crises. It is only in the last "Other factors" question where ToC specific details can be addressed e.g., relating to assumptions.
- Internal coherence of a country programme is a relevant concern and can be limited by "siloed" thinking. One indicator of this risk is where the ToC shows each activity in a separate causal pathway to the overall objective, as opposed to illustrating the multiple interconnections between these branches. The existence and workings (or not) of these cross linkages should be of particular concern to evaluation teams when assessing internal coherence.
- Contribution Analysis and Process Tracing were referred to as a means of doing theory-based evaluation in some

CSPEs and DEs. Other evaluation reports made references to Contribution Analysis as a means of doing theory-based evaluation but with limited further explanation of what that meant. To deal with this, ToRs could ask that naming of intended methodologies should be accompanied by references to the source documents that are expected to inform the use of this method, thus providing some form of accountability of intentions. Alternatively, linking subquestions to the methods that will be used to address them could provide more explicit connections. This is especially important with the use of terms like Realist Evaluations and Contribution Analysis which are arguably best described as approaches rather than methods.

- Not all evaluations took the opportunity to mix deductive hypothesis testing and inductive pattern finding, by attending to policy-as-theory and policy-as-practice. All theories are partial views of the world , so alternate views also need to be explored, albeit within the confines of the evaluation team's resource limitations. This happens to some extent with more general questions about the influence of external factors and drivers of change.

Good practices to be leveraged / explored

Elements of theory-based evaluation found in various evaluations examined should be leveraged across all evaluations using a theory-based approach. For instance:

- A reconstructed theory of change for the evaluand that outlines the outcome pathways and associated assumptions, with outcome pathways showing that outputs can contribute to multiple outcomes.
- The evaluation examines the logic model and testing the causal linkages presented in the reconstructed ToC through Contribution Analysis or other theory-based methods.
- The assumptions underlying the theory of change are assessed. A useful distinction could be made between context assumptions and causal assumptions (the latter about the connections between the elements of the ToC).
- These assumptions are captured in the evaluation matrix, which integrates the hypotheses formulated within the framework of the ToC.
- A comprehensive ToC with good narrative before the diagram.

4. Evaluability assessments and linkages with evaluation design

How could the linkages between the evaluability assessment, the theory of change and the selection of methods and data collection tools be further strengthened?

Overview of current practice across evaluations

As found in the 2022 Review, all evaluations examined included an evaluability assessment in the inception phase, building on the light assessment undertaken in the ToR. Most evaluations clearly indicated the scope of the Evaluability Assessment with the main focus on data availability, in line with the Inception Report guidance. Although in some cases that was done very thoroughly, data availability could be seen as only one of four facets of evaluability, namely:[iv] (a) data availability, (b) design (theory of change), (c) stakeholder demand, and (d) institutional and physical context.

Systemic challenges

Evaluability Assessments generally lacked explicit consideration of other stakeholder areas of interest and whether they could be reflected in the evaluation questions. In reality, the questions were almost wholly determined by the Evaluation Matrix template. Only in a few instances was there evidence of local stakeholders' interests and questions modifying the evaluation team's line of inquiry. Across CSPEs and DEs, context related issues such as physical, institutional and political constraints were often not examined in the evaluability assessments.

Other common weaknesses identified in the 2022 EMAP report were also found in 2023. For example, not all evaluations were clear about how the evaluability assessment informed the evaluation design at the inception report stage.

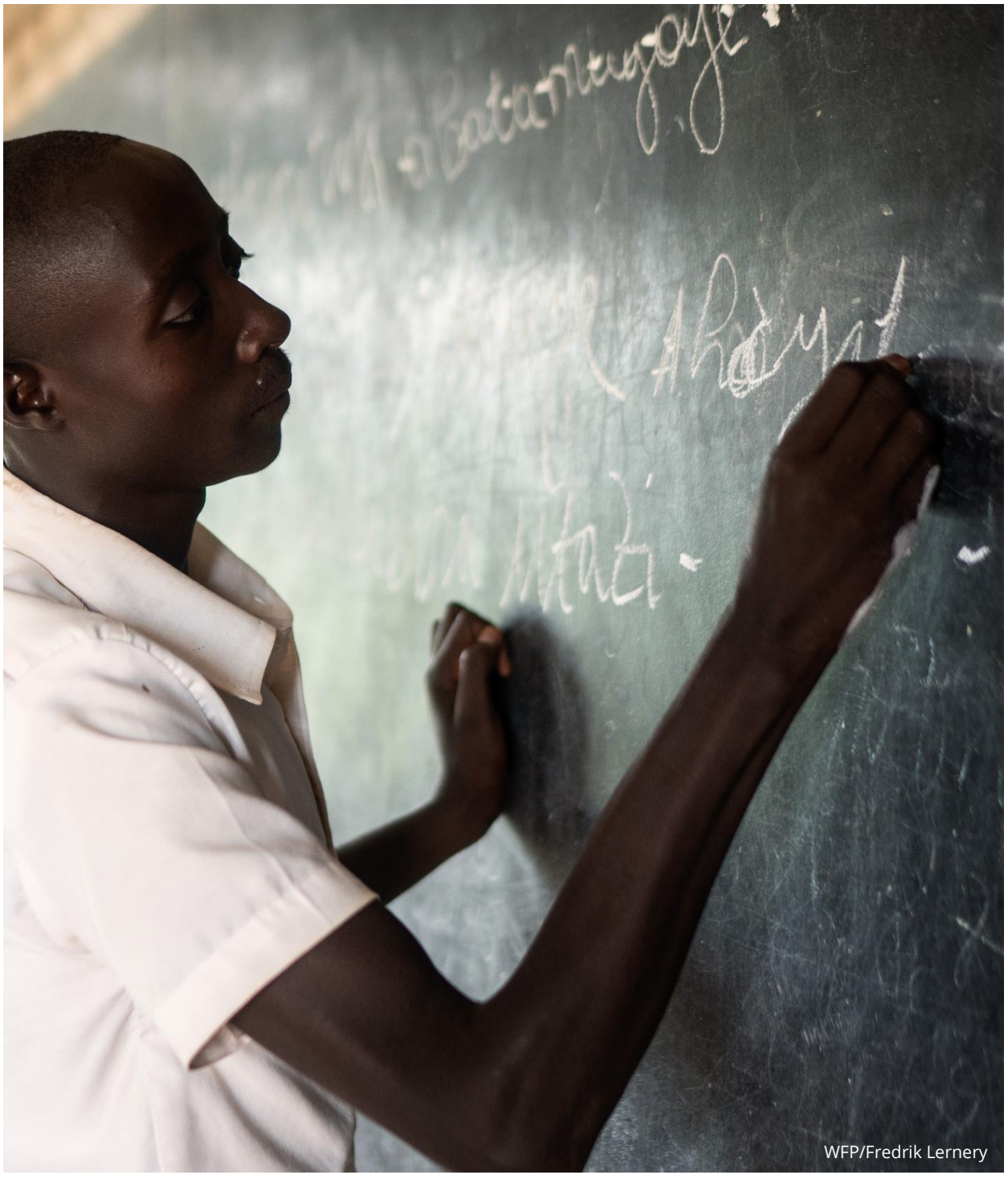
Moreover, there was limited use of the evaluation matrix as an instrument to clarify the evaluation design and, where appropriate, to show how assumptions underlying the evaluand's theory of change would be tested. Equally, not all the data sources identified in the evaluation matrix were addressed in evaluability assessments.

Good practices to be leveraged / explored

Evaluation reports displayed examples of good practice at all stages of evaluation preparation, design, and implementation, as exemplified in some evaluations:

- Inception reports described evaluability challenges and how the proposed methodology would manage them (e.g. South Sudan CSPE)
- Several evaluations included tables that matched evaluability challenges to proposed mitigating actions (e.g. Tanzania CSPE; South Sudan CSPE)

- The Algeria CSPE took a notably broad view of evaluability included examining the Strategic Outcomes, data availability, political sensitivity regarding beneficiaries, sampling, access to information outside the public domain, respondent accessibility and availability, team access to camps, and Covid-19 related issues.
- The Tajikistan CSPE Evaluability Assessment recognized the specific challenge of evaluating capacity strengthening; the evaluation report recommendations included the need to address issues with corporate indicators for assessing capacity development.
- The Peru DE used a red-amber-green rating to assess evaluability for each evaluation question, thus strengthening the evaluation matrix.



WFP/Fredrik Lernery

5. Triangulation, clarity, and transparency

How could triangulation be further strengthened to ensure clarity and transparency across data sources, data collection and analysis methods?

Overview of current practice across evaluations

Generally, the inception reports yielded comprehensive information on data sources and data collection tools that the evaluators intended to use, and on how they would triangulate sources and tools. There were some excellent examples of systematic, transparent triangulation of findings from different sources and instruments well-presented in the evaluation report.[v] Some ERs also explained how they combined analysis methods and evaluators' perspectives. But usually, an explanation of the strength of triangulation, or of what happened when different types of data did not agree, was missing[vi].

Overall, it was not easy to assess the strength of the evidence presented in evaluation reports. Information on analysis frameworks and methods was often patchy, with the evaluation reports rarely explaining whether and how a specific method – e.g., Process Tracing, Contribution Analysis or Outcome Mapping – had been fully applied. Across the board, there was no information on piloting and testing of qualitative research instruments. Reports on standardized surveys undertaken by evaluation teams often lacked basic information on population size, the characteristics of the sample, response, and completion rates.

Information on limitations often focused on context issues, e.g., travel and contact restrictions, and missing or poor-quality monitoring data. Those challenges were often described without explaining the implications for the findings. The evaluation reports rarely discussed issues linked to their own methodological choices, e.g., sampling or selection of respondents, completion rates in surveys, and light-touch use of certain approaches and methods. Correspondingly, the extent to which data validity and reliability was discussed varied between evaluations. Mandatory evaluation matrices provided a rough idea of triangulation strategies but were often vague regarding the sources of information and frequently stretched across more than five pages, making them difficult to use.

When presenting their findings, several evaluation reports mixed description, raw data, evidence from the data, and corresponding findings, hiding the evaluators' judgment.

Systemic challenges

Gaps in WFP project monitoring data were a frequent problem across evaluations. For instance, several evaluations teams pointed out that reporting on performance against relevant indicators was inconsistent and unclear. Gender disaggregated data and generally, data capturing effects for different population groups, were frequently missing. That made it difficult for evaluation teams to assess differential effects and the project's contribution to gender equality and broader equity.

In some DEs with a stated institutional learning purpose, the evaluation questions were mainly normative. Correspondingly, the evaluation reports focused on assessing project performance while neglecting important learning questions. They failed to collect appropriate data for learning, e.g., about unintended effects and how they came about. WFP templates used in evaluations are helpful, but they should be adapted to each evaluation so that they serve the purpose of the evaluation and only include necessary questions.


WFP/Antoine Vallas

Good practices to be leveraged / explored

Some evaluation reports displayed examples of good practice at all stages of evaluation preparation, design, and implementation, explaining why and how the planned methodology was adapted during the evaluation. For instance, the Tanzania CSPE used a table for that purpose and included lessons for future evaluations – a practice that could be encouraged in future evaluations.

There are always limits to data validity and reliability in any evaluation, as evaluation is not held to the higher standards of scientific research. WFP should encourage evaluation teams to be more open about these limits – not only considering gaps in secondary data, but also about primary data collected during the evaluation.

Some evaluation reports adopted a clear structure when presenting findings: each sub-section the findings chapter started with a clear statement of the finding and continued with a presentation of related evidence. Such practice enhances transparency and force evaluation teams to make clear evaluative judgements. The use of plain English in evaluation reports can also make it easier to detect the findings, which some reports have hidden behind jargon, and enhance evaluation use.

6. Lessons to strengthen WFP's evaluation function

Systemic considerations

It is recommended to **invest more time** into ToR development and dialogue between WFP and evaluation teams throughout all phases of the evaluation, to ensure:

- **ToR and evaluation questions or sub-questions are tailored to the project** or programme, to prioritised WFP information needs, and to the resources available for the evaluations.
- **WFP and evaluation teams have a shared, documented understanding** of the evaluation questions and definitions of key concepts for the evaluation early in the inception phase, so that the evaluation teams can propose an appropriate evaluation design in its draft inception report.
- **Evaluability is carefully assessed using a broad understanding of the concept and the evaluation design** is based on findings from the evaluability assessment.
- **Data collection in the field is well-prepared**, with appropriate, transparent sampling/selection processes, pretested data collection tools, and risk management strategies.
- **Communication between WFP and evaluators** continues in data collection and analysis phases so that emerging issues can be identified and dealt with.

In DEs, some WFP donors have stringent evaluation requirements, which can make it difficult to tailor the evaluation design to the specific project and to WFP's information needs. It is recommended to **raise donors' awareness** for the trade-offs between standard methodologies and maximum coverage of questions on one hand, and project-specific depth of analysis on the other hand.

Lack of good quality monitoring data can be an obstacle to effective evaluation. Evaluations report limitations caused by (i) scarce monitoring data (e.g., performance data, gender disaggregated data) and (ii) inconsistent and unclear reporting on performance against relevant indicators. Addressing data limitations in evaluability assessments and evaluation reports, possibly with suggestions for future improvement, is crucial to enhancing the evaluations rigor and depth. Equally, OEV could build on the recent synthesis of lessons related to monitoring and regularly assess the evidence for discussing with the relevant units responsible for monitoring in WFP.

Areas which don't require so much attention / to do less of

WFP guidance and templates are comprehensive and generally of good quality. Templates are particularly important, as evaluation commissioners and evaluation teams will not necessarily engage with the full guidance documents. **In addition to refining (but not necessarily expanding) written guidance for evaluation managers and evaluation teams, it may be helpful to reinforce direct dialogue** between evaluation managers and evaluation teams, as outlined in the section above.

In some cases, evaluation managers provide an existing evaluation report as an example to guide evaluation teams. That practice should be discontinued, as each report has its strengths and weaknesses, and **each evaluation should be tailored to its specific purpose and context. Rather, OEV should continue to extract good practices and share these among evaluation managers.**

Mixing qualitative and quantitative methods can make it easier to triangulate data collection and analysis instruments. But in some cases it may not be necessary to incorporate surveys, which can be costly and a burden on participants. Instead, as demonstrated in some reviewed CSPE, evaluation teams can use data and analyses from previous surveys or opportunistically add relevant questions to surveys that are already planned. Conversely, where the sole purpose of a DE is to assess project progress against a standardized baseline survey, a follow-up survey might suffice.

Likewise, theory-based approaches can be an excellent way to assess efficiency and likely impact, but they are not necessarily the best choice in all evaluations. Therefore, even though (re-) constructing the evaluand's **theory of change** is often an important step in process, this should not be routine requirement.

Key areas for attention to increase quality

There is a trade-off between the breadth of evaluation questions and the depth of analysis. Evaluations that are expected to contribute to institutional learning should aim for depth. That is possible with a manageable set of 3-5 main evaluation questions addressing only OECD-DAC criteria of priority importance in that evaluation. This approach is found in the standard set of questions established in guidelines for CSPEs and PEs.

Evaluation questions need to be designed to serve the purpose of the evaluation; the evaluation methodology needs to be developed to answer those questions, and **inception reports need to explain why certain methods and tools are proposed**. This could be done in the form of a reader-friendly research plan or a simple matrix that sets out the questions, sub-questions, and lines of enquiry in the inception report. On that basis, an evaluation matrix with details on methods and data sources can be prepared. All the evaluation reports should include a section **explaining departures from the ToR, and from the methodology proposed in the inception report.**

Gender/equity/inclusion aspects and evaluations are still very limited. While most reviewed evaluations have included at least basic information on gender inequality and some information on women's representation, full gender analysis is rare. Very few reports discuss aspects on broader equity and inclusion. A first step towards ensuring evaluations consider differential effects across population groups (as defined in the OECD-DAC effectiveness criterion) could be a requirement to deconstruct or **disaggregate the "affected population/communities" category** in the evaluation stakeholder mapping. Instead, evaluation teams could be asked to identify marginalised groups within that category and explain how they would be included in the evaluation.

Key areas for attention to increase utility

As pointed out above, **evaluation questions and evaluation design need to be in line with the purpose of the evaluation**. For instance, in formative evaluation, typical questions would be, "what works and what doesn't? Where are opportunities for improvement?" Where institutional learning is the main purpose of the evaluation, classical questions are: "What has worked, for whom, in what ways and under what conditions? What principles can be extracted across results to inform practices and models in new settings?" These examples are by Michael Q, Patton, whose work on facilitating evaluation offers excellent inspiration.[vii]

Likewise, more **continuous communication** between the evaluation manager, the evaluation team, and possibly other evaluation users will ensure the evaluation remains focused on the questions and fields can feed into decisions that will be made on the basis of the evaluation.

Finally, **more evaluation reports could be designed in a more user-friendly manner** – with a well-structured, unambiguous presentation of findings, clear, possibly plain language, and short, to-the-point summaries. In this respect, the 1-page infographics that are routinely produced for CSPEs may also be useful to some stakeholders.

WFP/Nelson Pacheco

Annex 1: Short biographies of members of the EMAP

Khalil Bitar	• +13 years of experience • Specialized in evaluation in countries affected by fragility, conflict, and violence; equity and social justice issues in evaluation; youth empowerment evaluation capacity strengthening, and transformative evaluation practices
Paul Knox Clarke	• +25 years of experience • Extensive experience in strategy development, organizational structures and the international humanitarian sector
Rick Davies	• +30 years of experience • Specific expertise on evaluation participatory approaches, social network analysis, theories of change, qualitative comparative analysis, evaluability assessments, most significant change
Michaela Raab	• +30 years of experience • Specific expertise on theories of change, evaluations of (portfolios of) complex interventions, human rights-based and gender-responsive evaluation and strategy development, qualitative comparative analysis
Michael Reynolds	• +30 years of experience • Strong experience with managing and conducting country programme evaluations and strategic evaluations
Patricia Rogers	• +30 years of experience • Expert in both quantitative and qualitative data for evaluation (e.g., cost benefit analysis, appreciate enquiry), and theory-based evaluation

Annex 2: Evaluation documents reviewed by the EMAP

Name of evaluation	Report	Reviewer
Evaluación final conjunta de piloto de protección social reactiva a emergencias en Arauca, Colombia, 2020-2021	ToR, IR, ER	Michaela Raab and Khalil Bitar
Evaluación del Efecto Estratégico 1 hacia los objetivos Hambre Cero a través de la abogacía, comunicación y movilización, del Plan Estratégico de País-Perú, 2017–2021	ToR, IR, ER	
Thematic Evaluation of Supply Chain outcomes in the Food System in Eastern Africa, 2016-2021	ToR, IR, ER	
Innovative Pilot Evaluation of Aflatoxin Reduction in the Rwanda Maize Value Chain, October to December 2021	ToR, IR, ER	
Joint Evaluation of the SADC Regional Vulnerability Assessment and Analysis (RVAA) programme (2017- 2022)	ToR, IR, ER	
Evaluation of the Asset Creation and Public Works Activities in Lesotho, 2015-2019	ToR, IR, ER	
Evaluation of R4 Rural Resilience Initiative in Masvingo and Rushinga Districts in Zimbabwe, 2018–2021	ToR, IR, ER	
Evaluation thématique des activités de renforcement des capacités institutionnelles en Guinée, 2019-2021	ToR, IR, ER	
Formative Evaluation of WFP Livelihoods Activities in Northeast Nigeria, 2018-2020	ToR, IR, ER	
Evaluation of WFP's Livelihood Activities in Türkiye, 2020-2022	ToR, IR, ER	
Evaluation of the First 1000 Days Programme in Egypt, 2017-2021	ToR, IR, ER	
Endline evaluation of USDA McGovern-Dole International Food for Education and Child Nutrition Programme in Nepal, 2017-2020	ToR, IR, ER	
Evaluation of WFP's support to smallholder farmers and expanded portfolio across the agriculture value chain in Bhutan, 2019-2021	ToR, IR, ER	
Synthesis of Evaluation Series on Emergency School Feeding in the Democratic Republic of Congo, Lebanon, Niger and Syria, 2015-2019	ToR, IR, ER	

Name of evaluation	Report	Reviewer
Algeria	ToR, IR, ER	Rick Davies and Mike Reynolds
Chad	ToR, IR, ER	
Kyrgyzstan	ToR, IR, ER	
Mauritania	ToR, IR, ER	
Palestine	ToR, IR, ER	
Pakistan	ToR, IR, ER	
Peru	ToR, IR, ER	
South Sudan	ToR, IR, ER	
Tajikistan	ToR, IR, ER	
Tanzania	ToR, IR, ER	
PE Country Strategic Plan Policy	Evaluation Report	Rick Davies
PE Resilience	Evaluation Report	
PE DRR / Climate Change	Evaluation Report	
PE Emergency Preparedness and Response	ToR	Paul Knox Clarke
PE Environmental Policy	ToR	Mike Reynolds
CEE Myanmar		Paul Knox Clarke
CEE Sahel Phase I	Final report	Patricia Rogers
CEE Sahel Phase II	Inception Report	Michaela Raab
SE PSEA	Concept Note, IR	Michaela Raab
SE Mid-Term Evaluation of the WFP Strategic Plan	Concept Note, IR	Mike Reynolds
SE refugees, IDPs, migrants	Concept Note, IR	Paul Knox Clarke

Annex 3: Selection of evaluations for review by the EMAP

Type of evaluation	Number of evaluations reviewed	Selection criteria
Decentralized evaluations	14	A sample of 1/3 of the reports, ensuring broad regional representation and based on the rationale for selection provided by the regional evaluation unit to focus on those with specific methodological aspects
Country strategic plan evaluations	10	A sample of 10 evaluations was selected from the 20 CSPEs completed in 2022. The purposive sample was based on evaluations that had two positive features: (a) covered a bigger proportion of the 11 topic areas of interest contained in the CSPE report feedback form and (b) covered topic areas in a way that was especially detailed and sometimes different from apparently more routine approaches.
Policy evaluations	5	All evaluation started in 2023 (including some started in 2022 and continuing into 2023)
Corporate emergency evaluations	3	All evaluation started in 2023 (including some started in 2022 and continuing into 2023)
Strategic evaluations	3	All evaluation started in 2023 (including some started in 2022 and continuing into 2023)

Endnotes

[i] The WFP technical note (TN) on evaluation criteria includes all DAC criteria and adds the ALNAP criteria of appropriateness, connectedness, and coverage in humanitarian contexts. It allows for excluding criteria but requires explanation if this is done.

[ii] This could be the case, for example, for an evaluation which uses only survey data or only data from interviews.

[iii] But in some CSPEs there were Logical Frameworks, which can be considered as a type of Theory of Change (minus the detailing of causal links between events). Additionally, although not requested, it should be noted that there did not appear to be any examples of the ToCs being subsequently revised in the light of the evaluation findings.

[iv] Austrian Development Agency. 2022. *Evaluability Assessments in Austrian development cooperation*. Guidance Document.

[v] Conversely, there was one example of a DE that appeared to use internal sources only.

[vi] An "Evidence Table" where an assessment is made for each sub-question in the evaluation matrix is a useful tool to identify where evidence is adequate or not.

[vii] Michael Q. Patton (2018) *Evaluation in Practice Series 2: Facilitating Evaluation*. Thousand Oaks: Sage. Pages 146-147.