



International
Labour
Organization

Evaluation
Office

► *i*-eval THINK Piece, No. 28

Results and reflections from a quality appraisal of 2023-24 ILO evaluation reports, with a focus on the value of participatory approaches

October 2025



The data and findings presented in the Think Piece are summarized from a separate external study prepared by DeftEdge LLC and was developed by Ricardo Furman (EVAL).

This report has not been professionally edited.

► Contents

1. INTRODUCTION	5
2. FINDINGS	8
3. CONCLUSIONS	15





1. Introduction

Background

As part of the annual quality assurance process of the International Labour Organization's (ILO's) Evaluation Office (EVAL),¹ Universal Management Group Limited and DeftEdge (hereinafter "the review team") conducted the quality appraisal (QA) of 74 independent evaluations of development cooperation projects that were conducted in 2023-24. A total of 72 evaluations were appraised by the first firm, and 2 by the second firm². The independent evaluation reports appraised include a mix of final and mid-term evaluations, in English, French, and Spanish, covering all regions where the ILO operates, and involving several ILO departments and offices.

► EVAL has conducted regular ex-post quality assurance of decentralized project evaluations since 2015. In 2020, EVAL transitioned from conducting quality assurance on a sample of independent evaluations to implementing a rolling ex-post quality assurance process that covers all independent evaluations. Once evaluations are reviewed and approved by EVAL, the external review team is tasked to appraise them on an ongoing basis, allowing EVAL to rapidly identify and act upon emerging quality issues. Results are disseminated to the ILO's evaluation network to encourage learning in a targeted manner.

The consolidated analysis and results presented in this Think Piece have been developed by DeftEdge. The main purpose of this report is to provide a cumulative analysis of evaluations that were conducted in the reporting period and to assess trends and comparisons with previous quality appraisals. This report informed the ILO's latest [Annual evaluation report 2024-5](#), released in October 2025.

This Think Piece summarizes the findings, conclusions, and recommendations of the QAs of the analysed timeframe. In addition to the introduction, it contains the following sections:

- **Section 2** presents the main findings of the quality appraisal; and
- **Section 3** presents the main conclusions of the quality appraisal.

1 This is one in a series of external quality appraisals of independent evaluation reports. A total of 12 quality appraisals has been conducted since 2014. The results have been described in various [Think Pieces](#).

2 Due to logistic adjustments in the QA process, 11 evaluation reports conducted in 2023 were not able to be integrated in the analysis corresponding to that year and therefore were part of the consolidated analysis of 2024. The 11 evaluations from 2023 cover 15 per cent of the 74 evaluations analysed.

Overview and implementation of the quality appraisal tool

The ILO's quality appraisal tool examines four dimensions, as described below, that are based on [EVAL's quality appraisal checklist](#) – allowing the review team to collect quantitative data on the quality of the ILO's evaluation reports.

First, the tool captures descriptive data on demographic variables for each evaluation report, such as the region, department, and year of completion. Data is then analyzed and aggregated to identify trends across variables.

Second, the tool requires the reviewers' team to rate the quality of the evaluation reports' content according to 58 different items that are grouped across the evaluation report's ten standard sections.

Third, the tool reviews whether key components, as per EVAL's quality appraisal checklist, are included in the report, using a two-point scale (absent-present).

Fourth, the tool which is based on the United Nations System-Wide Action Plan (UN-SWAP) assesses the integration of gender in alignment with the [Guidance on Integrating Human Rights and Gender Equality in Evaluation](#).

The ex-post external quality appraisal of evaluation reports is a four-step process, involving two reviewers per evaluation to ensure inter-rater reliability. Once all reports of a given reporting period are appraised, quantitative ratings and qualitative information justifying the rating are aggregated in an Excel sheet, and overall scores are calculated for the quality, comprehensiveness, and UN-SWAP dimensions. Aggregated scores and individual ratings are analysed using quantitative and qualitative methods.

The review team adopts a concurrent triangulation design as it simultaneously generates and analyse the quantitative and qualitative data. This method is used to cross-validate findings and provide plausible explanatory factors for the tendencies that are observed.

The quality appraisal process is based on the [ILO's evaluation policy, strategy](#), and [guidelines](#) (such as checklists), and is aligned with [United Nations Evaluation Group \(UNEG\) evaluation norms and standards](#). In addition, the review team is committed to undertaking a gender-responsive quality appraisal by assessing whether gender is adequately mainstreamed across the evaluation reports appraised.



2. Findings

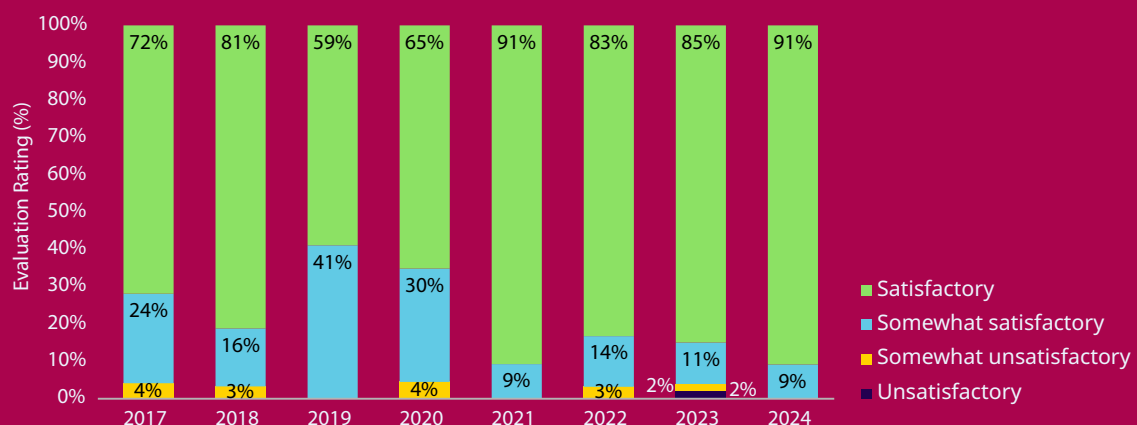
Overall quality score

All of the appraised independent evaluation reports assessed met the required quality standards ("somewhat satisfactory - SS" or "satisfactory -S" median rating) in 2024 (figure1). While 9 per cent of reports obtained a "SS" rating, 91 per cent obtained a "S" rating. None of them received an overall assessment of "unsatisfactory -US", "highly unsatisfactory -HU" or "highly satisfactory- HS".

Overall, the quality of independent evaluation reports has remained satisfactory since 2016 (overall rating of 5 out of 6). The dispersion of scores has remained low since 2015, suggesting a certain homogeneity in the quality of reports over the years.

In the reported period, the quality of independent evaluation reports remained consistent across the departments and offices represented in the sample, all departments and offices obtained a median score equivalent to a "satisfactory" rating (5 out of 6) for the evaluations overseen by evaluation managers on behalf of EVAL for the evaluated period.

► Figure 1. Overall median score and evolution of ratings per year³



Regarding the independent evaluation reports' components, the analysis found that quality was quite consistent across the departments and regional and country offices represented in the sample, with eight out of ten components obtaining a "satisfactory" rating and the other two "somewhat satisfactory" (median score of 5 out of 6) for the evaluations.

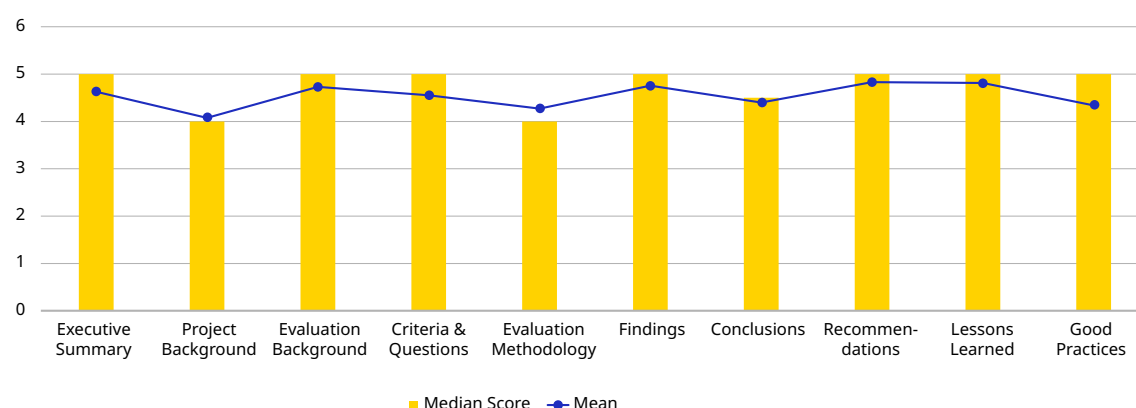
³ Results for 2015 and 2016 are taken from the quality appraisal conducted in 2017 by Artival. Results for 2017 and 2018 are taken from the QA conducted in 2019 by Universalia. Results for 2019, 2020, 2021, 2022 and 2023 include all the evaluation reports that were appraised through real-time QA by Universalia. Results for 2024 include 11 evaluation reports from 2023 and 63 from 2024 (72 appraisals conducted by Universalia and 2 by DeftEdge).

Quality of individual components

This section looks at the component-level performance of the portfolio, including comparisons by year, and then provides an analysis of the ratings for the items within each component.

As per figure 2, seven of the 10 components had a 'Satisfactory' rating in 2024 with median scores of 5 (out of 6). Project Background and Evaluation Methods were rated as 'Somewhat Satisfactory' with median scores of 4. Conclusions landed between these two categories with a median score of 4.5.

► **Figure 2. Components' scores for 2024**



Comprehensiveness of evaluation reports

The QA tool includes a section on whether evaluation reports follow the expected structure. A binary scale is used to quantify the presence or absence of specific components.

The 2024 portfolio mostly included EVAL's mandatory components for evaluation reports. As indicated in table 1, the average rating in 2024 was 94.4 per cent - similar to 2023. Of the two persistent weaknesses, there continues to be a significant number of reports that do not include a 'list of tables, figures and charts' (over one third of evaluations), however more reports included all expected annexes compared to 2023.

► **Table 1. QA results on comprehensiveness of independent project evaluation reports (2015-24)**

ITEMS	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Title page using EVAL's template	67%	73%	92%	91%	95%	100%	100%	100%	100%	97%
Table of contents	98%	95%	96%	100%	100%	100%	100%	99%	98%	99%
List of tables, figures and charts	33%	38%	36%	44%	54%	39%	62%	55%	61%	61%
List of annexes	79%	95%	92%	88%	95%	98%	96%	97%	87%	92%
List of acronyms or abbreviations	100%	92%	100%	94%	100%	98%	100%	96%	100%	97%
Executive summary	100%	97%	100%	97%	97%	96%	98%	99%	100%	100%
Project background	100%	97%	100%	100%	100%	100%	100%	100%	100%	99%
Evaluation background	100%	100%	100%	100%	100%	98%	100%	100%	100%	99%

ITEMS	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Criteria and questions	72%	84%	72%	75%	82%	94%	100%	97%	96%	96%
Methodology	100%	97%	100%	100%	100%	100%	100%	100%	100%	100%
Findings	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Conclusions	95%	100%	92%	94%	98%	100%	100%	96%	96%	100%
Lessons learned	98%	100%	100%	100%	97%	98%	98%	100%	100%	100%
Emerging good practices	81%	89%	84%	84%	85%	85%	96%	97%	94%	95%
Recommendations	100%	100%	100%	100%	100%	100%	100%	100%	100%	99%
Annexes	70%	60%	76%	66%	67%	59%	71%	76%	78%	85%
Average	87%	89%	90%	89%	92%	91%	95%	94%	94%	95%

UN-SWAP

DeftEdge also analysed the extent to which the ILO met UN-SWAP requirements in the reported period. The ILO is one of 69 organizations that are mandated to report on the United Nations System-Wide Action Plan on Gender Equality and the Empowerment of Women (UN-SWAP-GEEW).

Reports are submitted on an annual basis using the UNEG-endorsed Technical Note and related scorecard to report against the Evaluation Performance Indicator (EPI).⁴ In order to comply with this requirement on an annual basis, the ILO conducts external quality appraisals (best practice) instead of self-assessments to rate the EPI in the UN-SWAP-GEEW scorecard. The goal is for all UN system entities to “meet requirements” related to the EPI in terms of integrating the GEEW in their respective evaluations.

Evaluation reports were rated on the extent to which they met the three UN-SWAP criteria for Gender Equality and the Empowerment of Women (GEEW), using the rubric set out in the SWAP scorecard. A meta-score is then obtained by aggregating the criteria scores for all evaluations.

The meta-score calculated for the 2024 portfolio was 6.0 (median) and 5.81 (mean) which is categorized as “Approaching Requirements”. The full distribution of scores and comparison with 2023 results are presented below in table 2. Precisely half of 2024 reports (50 per cent) scored in the range of ‘5’ to ‘7’. Notably, 32 per cent met requirements, continuing the trend of year-on-year improvement. ***The percentage of reports meeting requirements has increased steadily from 25 per cent in 2021, to 27 per cent in 2022, to 28 per cent in 2023, and 32 per cent in 2024.***

► Table 2. Number and percentage of reports per point level

Categories	Misses requirements			Approaches requirements			Meets requirements		
Rating	1	2	3	4	5	6	7	8	9
N	2	3	1	1	21	22	15	8	1
%	2.7%	4.1%	1.4%	1.4%	28.4%	29.7%	20.3%	10.8%	1.4%
Total 2024	6 reports - 8%			44 reports - 59%			24 reports - 32%		
Total 2023	3 reports - 5%			36 reports - 67%			15 reports - 28%		

4 UNEG, UN-SWAP Evaluation Performance Indicator Technical Note, April 2018.

The performance of the portfolio for each of the three criterion is presented in table 3. The highest scores were seen for “G.2 Gender responsive methodology”, followed by “G.1 Integration of GEEW into scope, criteria and questions”. For both criteria, the majority of reports were satisfactorily or fully integrated. More shortcomings were seen in “G.3 Findings, conclusions, recommendations reflect a gender analysis”, where less than half of reports achieved a satisfactory rating and 9 per cent did not address gender issues.

► **Table 3 Percentage of reports' average scores per UN-SWAP criterion for 2024**

UN-SWAP Criteria	Not at all integrated (0)	Partially integrated (1)	Satisfactorily integrated (2)	Fully integrated (3)
G.1 GEEW is integrated in the evaluation scope of analysis and evaluation criteria and questions are designed in a way that ensures GEEW related data will be collected	3%	17%	57%	23%
G.2 A gender responsive methodology, methods and tools, and data analysis techniques are selected	1%	15%	49%	35%
G.3 The evaluation findings, conclusions, and recommendations reflect a gender analysis	9%	51%	35%	5%

Disability Inclusion

Disability inclusion in ILO evaluations was inconsistent. ***Almost half of the portfolio (45 per cent) did not adequately address the topic in their evaluation frameworks, with 38 per cent not having evaluation questions that directly referenced disability inclusion.*** However, appraisers also noted several positive examples where disability inclusion was well addressed, in some cases also within the contextual analysis and across multiple evaluation criteria.

Participatory practices

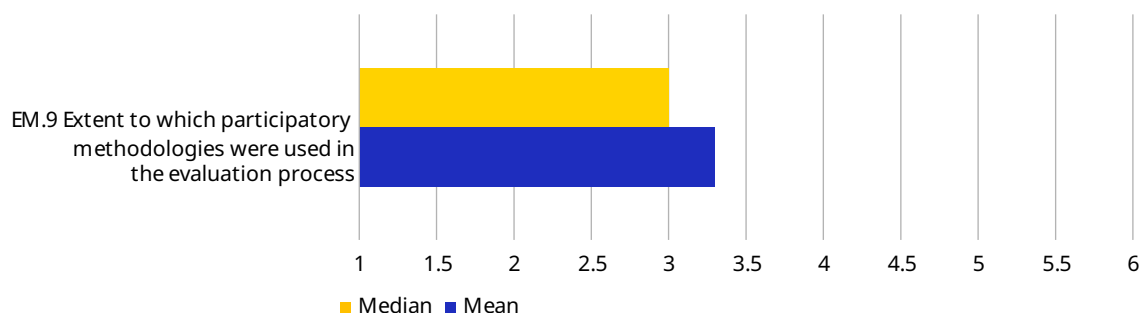
The ILO Results-Based Evaluation Strategy (2023-2025) emphasizes that evaluations should incorporate participatory approaches to improve ownership and accountability towards ultimate beneficiaries. To identify the prevalence of participatory practices and measure progress over time, in 2024 an item was added to the QA Tool to assess the extent to which participatory methodologies were used (item EM.9). The rubric for assigning this rating is shown in table 4. To achieve the highest rating of 6, there must be evidence in the evaluation report of multiple participatory practices ranging from stakeholders' mapping to a dissemination plan for sharing evaluation results with all categories of stakeholders.

► **Table 4. Assessment of use of participatory practices**

EM.9: Extent to which participatory methodologies were used in the evaluation process, including mapping of stakeholders, data collection techniques (e.g., focus groups, graphic participatory techniques), data analysis techniques (e.g., sense-making workshops with community members and right holders, and dissemination techniques (e.g., participatory workshops of discussion of preliminary findings in local languages, graphic presentations, etc.).

Rating scale and criteria					
1	2	3	4	5	6
No indication of participatory approaches used throughout the evaluation	Participatory approach used is limited to mapping the different categories of project stakeholders and their relationships	Data collection applies participatory techniques that cover most categories of mapped stakeholders	In addition to requirements for #3, a stakeholders' workshop held after data collection that includes participation of non-traditional stakeholders	In addition to requirements for #4, data analysis and reporting conducted in ways that can be understood by non-traditional stakeholders	In addition to requirements for #5, report includes a dissemination plan to share the report with all categories of stakeholders mapped at the inception phase

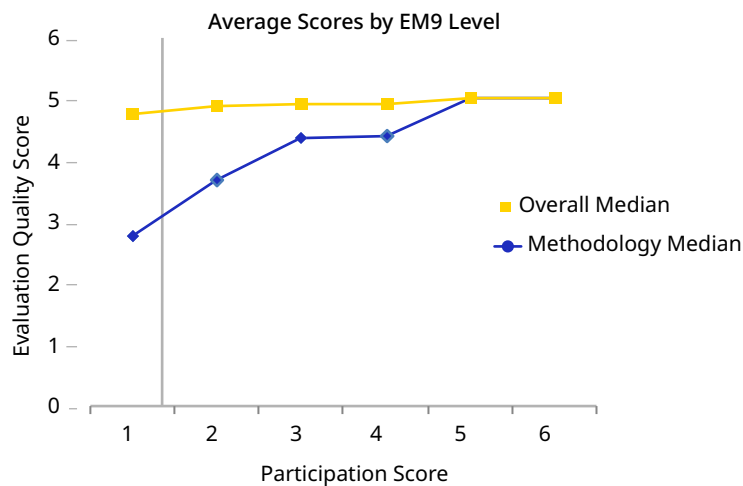
The degree to which 2024 evaluations managed to meet these expectations was mixed. Most evaluations explicitly incorporated some participatory methodologies, with appraisers noting a diversity of techniques including stakeholder workshops, focus groups, mapping exercises, validation sessions, and other feedback processes. However, the aggregated ratings for this criterion were relatively low compared with other assessed criteria. The median rating across the portfolio for EM.9 was of 3 out of a possible 6 (as shown in figure 3), less than half of the evaluations (39 per cent) received a rating of 4 or above, and only one achieved the highest score of 6. The application of participatory methodologies was often limited to a basic level of stakeholder mapping and data collection targeting most categories of the mapped stakeholders.

► **Figure 3. Ratings for use of participatory methodologies**

Analysis of the relationship between participatory processes and evaluation quality reveals important patterns for strengthening evaluation practice. **A strong positive correlation ($r=0.537$) was found between the extent of participatory methodologies used and the quality of evaluation methodology, thus confirming the rationale for emphasizing participatory processes.** Moderate positive correlations were observed between participatory processes and both lessons learned quality ($r=0.328$) and overall evaluation quality ($r=0.328$). However, the relationship with substantive outputs proved weaker, with negligible correlations found for findings quality ($r=0.081$) and recommendations quality ($r=0.089$), and only a weak correlation with conclusions quality ($r=0.157$).

As shown in figure 4, higher stakeholder participation correlated with better evaluation quality, with a particularly strong "threshold effect" at the highest levels. Figure 4 further illustrates this point; while 43 per cent of evaluations (32 out of 74) operated at moderate participation levels (EM.9 = 3) and achieved satisfactory quality scores for the overall report (with a mean of 4.8), those few evaluations that invested in high participatory processes (EM.9 = 5-6, representing only 12% or 9 of 74 evaluations) consistently achieved quality scores of 5 across recommendations, lessons learned, and overall quality.

► **Figure 4. Relationship between participation and overall evaluation quality**





3. Conclusions

Overall performance - The portfolio of 2024 evaluations continued to meet ILO standards for evaluation quality. The overall rating of “Satisfactory” was consistent with previous years. Also consistent was that no reports were rated as less than Satisfactory, but also that none attained a “Highly Satisfactory” rating even though good practices were evident. A notable improvement was the overall UNSWAP score for GEEW edging higher than it was before.

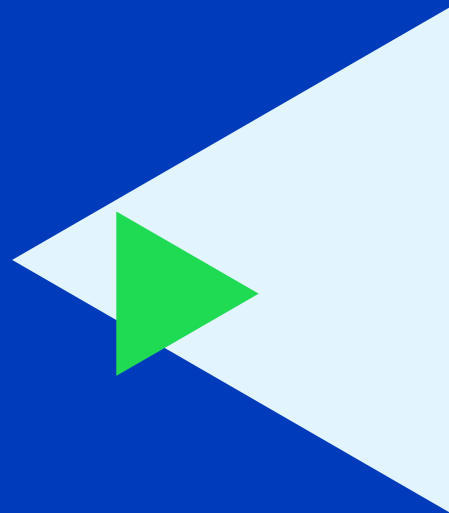
Main strengths and weaknesses - In addition to gender considerations, the stronger components across the 2024 portfolio were Evaluation Background, Findings, Recommendations and Lessons Learned. Weaker ones were Project Background and Evaluation Methods which were rated as ‘Somewhat Satisfactory’ having median scores of 4. Conclusions landed between these two categories with a median score of 4.5.

UN-SWAP performance - While most evaluations demonstrated satisfactory or strong integration of gender issues, gaps continued in explaining gender-responsive methodologies, providing intersectional analyses, addressing unintended effects of the intervention for those more likely to face situations of vulnerability, and carrying through GEEW-related findings to recommendations.

Participatory approach – *There is significant room to incorporate participatory methodologies in ILO evaluations.* The application of participatory methodologies was often limited to a basic level of stakeholder mapping and data collection. The median rating across the portfolio for this criterion was 3 out of 6 with few achieved a mark of 5. These last ones consistently achieved quality scores of 5 across recommendations, lessons learned, and overall quality, confirming the rationale for emphasizing participatory processes.

Key areas for improvement - Within Project Background, *the lowest ratings were seen for stakeholder mapping and description of context.* Within *Methodology*, *the key issues were inadequate explanations of sampling and analysis processes*, of how UNEG norms and standards were upheld (particularly adherence to ethical considerations).

A concerning trend is the number of reports that did not address ILO’s cross-cutting themes, with the exception of gender, in their evaluation frameworks. Most frequently absent was coverage of *disability inclusion, environmental sustainability, and SDGs*. Persistent issues in other components included unexpected outcomes not being addressed in findings, validity and reliability of conclusions not being explained, stakeholder consultation not being evident in the development of recommendations, and the applicability of lessons learned and good practices not being adequately apparent to other contexts. However, there was improvement seen for several of these issues in 2024.



ilo.org

International Labour Organization
Route des Morillons 4
1211 Geneva 22
Switzerland