



International
Labour
Organization

Evaluation
Office

► *i*-eval THINK Piece, No. 30

Results and reflections from a quality appraisal of 2025 ILO evaluation reports, with a focus on the value of participatory approaches

July 2026



The data and findings presented in the Think Piece are summarized from a separate external study prepared by DeftEdge LLC and managed by Ricardo Furman.

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 DeftEdge (hereinafter "the review team") conducted the quality appraisal (QA) of 50 independent evaluations of development cooperation projects that were conducted in 2025. 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 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, including a particular focus on participatory approaches as reflected in Section 2.6, which provides a more in-depth analysis of how participation was incorporated across the design, implementation, monitoring, and evaluation of interventions; 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 13 quality appraisals has been conducted since 2014. The results have been described in various [Think Pieces](#).

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](#) as well as on people living with disabilities in alignment with the [UN Disability Inclusion Strategy](#).

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 and UNDIS 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.



tps
TOH-SHI
PRINTING SERVICE

トホプリントサービス

2. Findings

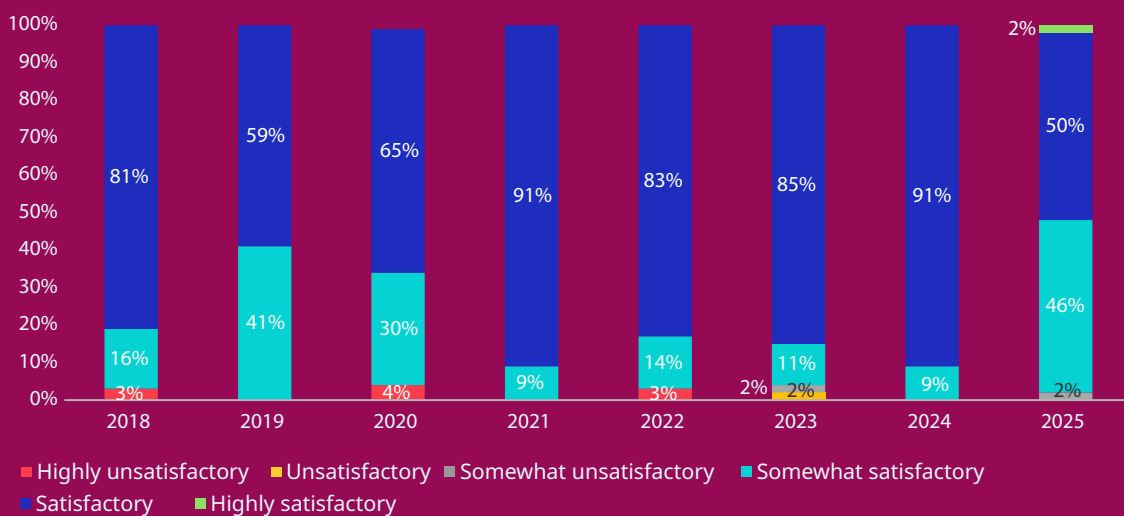
Overall quality score

The 2025 portfolio achieved an overall median score of 5 (on a 1–6 scale), consistent with the ILO's long-standing quality benchmark. However, the distribution of ratings within this threshold shifted: 52% of reports achieved a Satisfactory or above rating (50% Satisfactory, 2% Highly Satisfactory), with 46% Somewhat Satisfactory and 2% Somewhat Unsatisfactory (Figure 1).

Overall, the quality of independent evaluation reports has remained not less than somewhat 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.

Mean scores by region reveal that 2025 evaluations of projects located in Arab States achieved the highest mean overall score (4.62), followed by the Americas (4.47), Africa (4.40), Asia Pacific (4.22), Inter-regional (4.19), and Europe (3.99).

► **Figure 1. Evolution of ratings by year**



Quality of individual components

This section looks at the component-level performance of the portfolio. The analysis found that, as Figure 2 shows, evaluations received the highest scores for the Evaluation Background (median 5) and Criteria and Questions (median 6) components. Executive Summary, Project Background, Conclusions, and Recommendations were scored at a median of 4.5. Evaluation Methods, Findings, Lessons Learned and Good Practices each scored a median of 4 (Somewhat Satisfactory).

► **Figure 2. Components' scores for 2025**

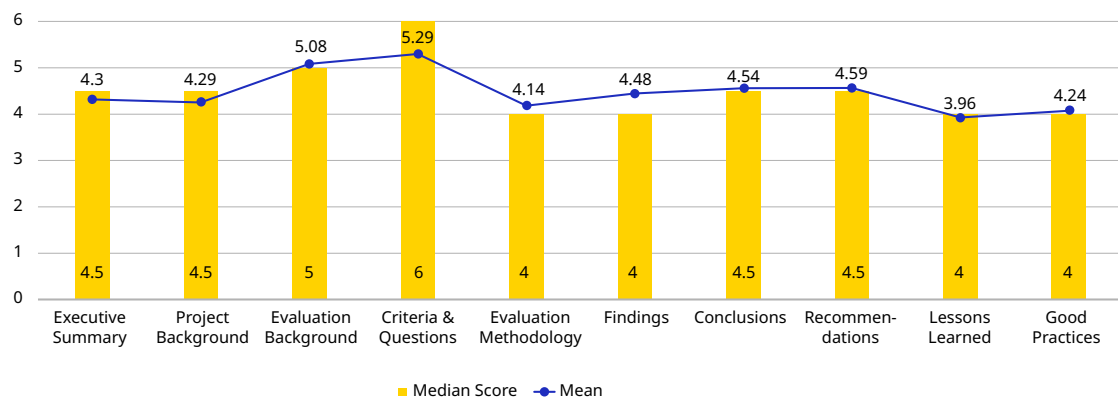
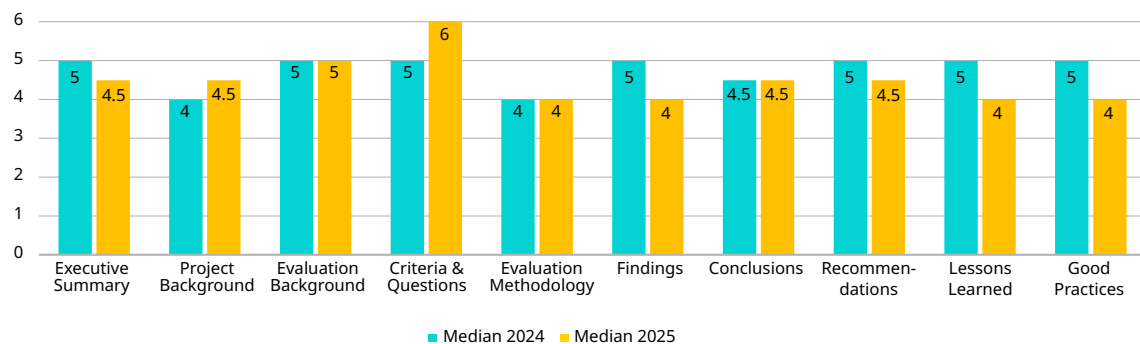


Figure 3 shows the evolution of component medians since the last report. Criteria and evaluation questions reached a median of 6 for the first time, continuing a multi-year upward trend. Evaluation Background has been consistently strong, maintaining a median of 5 since 2024. The most notable regression in 2025 is in executive summary and recommendations (5 → 4.5), and Findings, Lessons Learned and Good Practices (5 → 4).

► **Figure 3. Evolution of component medians since the last report**



Comprehensiveness of evaluation reports

The 2025 cohort achieved an average comprehensiveness rating of 93.3%, comparable to 94.4% in 2024. Nine core structural components were present in all 50 reports (Title Page, List of Acronyms, Executive Summary, Project Background, Evaluation Background, Methodology, Findings, Conclusions, and Recommendations), with Table of Contents (98%) and Criteria and Questions (94%) also performing strongly. The two persistent weaknesses were in supplementary components: inclusion of a List of Tables/Figures (68%) and Annexes (48%), which has declined noticeably from 85% in 2024 (Table 1).

► **Table 1. QA results on comprehensiveness of evaluation reports (2017-25)**

COMPONENT	2018	2019	2020	2021	2022	2023	2024	2025
Title page using EVAL's template	91%	95%	100%	100%	100%	100%	97%	100%
Table of contents	100%	100%	100%	100%	99%	98%	99%	98%
List of tables, figures and charts	44%	54%	39%	62%	55%	61%	61%	68%
List of annexes	88%	95%	98%	96%	97%	87%	92%	92%
List of acronyms or abbreviations	94%	100%	98%	100%	96%	100%	97%	100%
Executive summary	97%	97%	96%	98%	99%	100%	100%	100%
Project background	100%	100%	100%	100%	100%	100%	99%	100%
Evaluation background	100%	100%	98%	100%	100%	100%	99%	100%
Criteria and questions	75%	82%	94%	100%	97%	96%	96%	94%
Methodology	100%	100%	100%	100%	100%	100%	100%	100%
Findings	100%	100%	100%	100%	100%	100%	100%	100%
Conclusions	94%	98%	100%	100%	96%	96%	100%	100%
Lessons learned	100%	97%	98%	98%	100%	100%	100%	96%
Emerging good practices	84%	85%	85%	96%	97%	94%	95%	96%
Recommendations	100%	100%	100%	100%	100%	100%	99%	100%
Annexes	66%	67%	59%	71%	76%	78%	85%	48%
Average	89%	92%	91%	95%	94%	94%	95%	93.2%

One finding is particularly notable. Annex completeness dropped sharply to 48.0% in 2025 — the lowest rate on record and a significant regression from 85% in 2024. This is the most concerning comprehensiveness finding in this report. Missing annexes most commonly were the ToR, lists of interviewees, and document review lists, which are important for transparency and reproducibility of evaluation evidence. The inclusion of a list of tables, figures and charts also remained low at 68.0%, though this represents a marginal improvement from 61% in 2024.

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).² 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.

² UNEG, [UN-SWAP Evaluation Performance Indicator Technical Note](#), April 2018.

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 for the 2025 portfolio was 6.5 (median) and 6.48 (mean), categorised as 'Approaching Requirements'. This represents an improvement from the 2024 meta-score of 6.0 (median) and 5.81 (mean). The full distribution of scores and comparison with 2024 are presented below in Table 12. No report in the 2025 cohort falls into the "Misses requirement" band (score 0–3) — a notable improvement from 2024, when 8.1% of reports (6 of 74) missed the requirement. Exactly half of the 2025 cohort (25 of 50 reports, 50%) '*approaches the requirement*' (score 4–6), and the other half (25 of 50, 50%) '*meets the requirement*' (score 7–9, with at least one criterion rated "Fully integrated"). By comparison, in 2024, 59.5% of reports '*approached*' and 32.4% '*met*' the requirement.

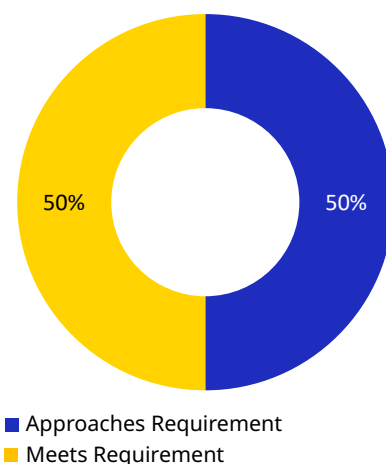


Table 2. Number and percent of reports per point level (2025 vs 2024)

Categories	Misses requirements			Approaches requirements			Meets requirements		
Score (points)	1	2	3	4	5	6	7	8	9
2025 - N	0	0	0	1	5	19	19	6	0
2025 - %	0.0%	0.0%	0.0%	2.0%	10.0%	38.0%	38.0%	12.0%	0.0%
2024 - N	2	3	1	1	21	22	15	8	1
2024 - %	2.7%	4.1%	1.4%	1.4%	28.4%	29.7%	20.3%	10.8%	1.4%

The portfolio median meta-score was 6.5 and the mean was 6.48, compared to a mean of 5.81 in 2024. Of the 50 evaluations, 50% (n=25) met requirements on the UN-SWAP scorecard, with the remaining 50% (n=25) approaching requirements.

► **Table 3. UN-SWAP Criterion Performance — 2025 Portfolio**

UN-SWAP Criteria	Not at all integrated (0)	Partially integrated (1)	Satisfactorily integrated (2)	Fully integrated (3)
G.1 GEEW in scope, criteria and questions	0%	2%	44%	54%
G.2 Gender-responsive methodology, methods and tools	0%	12%	72%	16%
G.3 Findings, conclusions and recommendations reflect gender analysis	0%	20%	68%	12%

Disability Inclusion

The evaluations were assessed for the extent to which they met expectations as per [UNDIS 2.0](#), and as with the UNSWAP scores for GEEW, this assessment is independent of the overall report ratings. The DIS results show that disability inclusion across the 2025 portfolio remains an area for development. The weakest dimension was disability analysis within findings, where two-thirds of reports showed no

engagement with disability and only 8% were rated fully satisfactory. Evaluation questions explicitly referenced disability in less than half of reports. Conclusions were comparatively the strongest area, with just over half rated "Mostly" or "Yes". Recommendations showed a more divided picture: nearly half of reports included no disability-related recommendations even when findings highlighted major gaps, while approximately a quarter did so fully. Evaluations typically scored higher for this component when issues around disability inclusion was addressed in an explicit manner instead of under the broader umbrella of vulnerable groups.

► **Table 4. Disability inclusion performance by criterion**

Criterion	Mean (0-3)	Median (0-3)
H.1 — Evaluation questions reference disability	1.34	1
H.2 — Disability analysis in findings	0.62	0
H.3 — Conclusions reflect disability	1.66	2
H.4 — Recommendations include disability	1.12	1
Overall	4.74	4

Participatory practices

The ILO Results-Based Evaluation Strategy (2023-2026) 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 5. 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 5. 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

Extent of participatory methodologies used) had a median of 4 in 2025, an improvement from the median of 3 in 2024. The score distribution shows that 40% of reports received a score of 3 (Somewhat Unsatisfactory), 42% received a score of 4 (Somewhat Satisfactory), 8% received a score of 5 (Satisfactory),

and one report (2%) achieved a score of 6 (Highly Satisfactory). Four reports (8%) received a score of 2 (Unsatisfactory), and no reports scored 1 (Figure 4).

Analysis of the relationship between participatory practices and evaluation quality reveals important patterns. A moderate positive correlation ($r \approx 0.48$) was found between EM.9 and overall evaluation quality, confirming that evaluations with more participatory approaches tend to score higher overall. A moderate positive correlation ($r \approx 0.376$) was also observed between EM.9 and the overall Evaluation Methods component score, indicating that participatory evaluations used more rigorous processes and/or documented their methodology more thoroughly.

► **Figure 4. Scores for Participatory Methods**

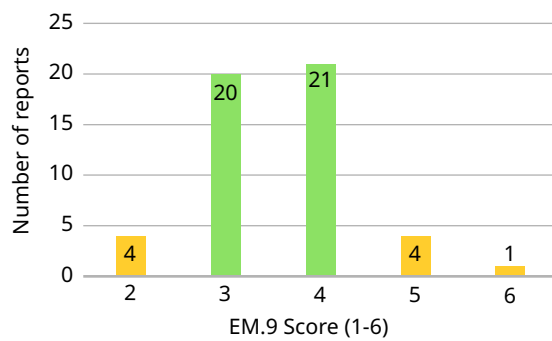


Table 6. Mean Quality Scores I

Participatory methods Level	Overall Quality Mean
2 (n=4)	3.99
3 (n=20)	4.28
4 (n=21)	4.43
5 (n=4)	4.57
6 (n=1)	5.05



3. Conclusions

Overall performance - The portfolio of 2025 evaluations continued to meet ILO's standards for evaluation quality, with an overall median score of 5 (Satisfactory) maintained for the ninth consecutive year. Unlike 2024, where ratings clustered at Satisfactory with none falling below that or reaching Highly Satisfactory, the 2025 cohort saw both ends of the scale represented: one report achieved a Highly Satisfactory rating and one fell to Somewhat Unsatisfactory, with just over half (52%) rated Satisfactory or above. A clear improvement was the UN-SWAP meta-score for GEEW with half of all evaluations now meeting requirements, up from around a third the previous year.

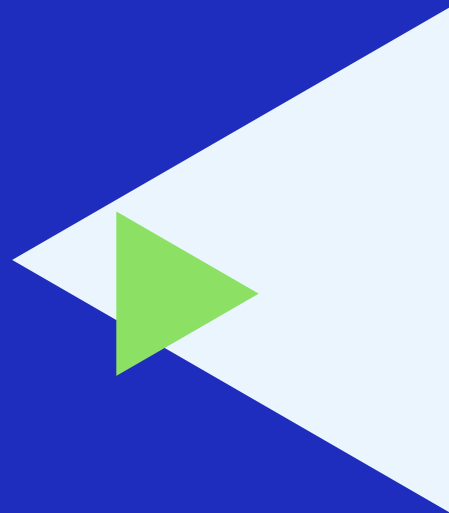
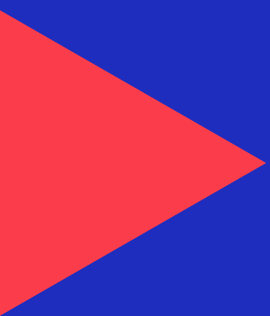
Main strengths and weaknesses - The stronger components across the 2025 portfolio were Cross-cutting Elements and Evaluation Background with the weaker components being Evaluation Methods, Findings, Lessons Learned and Emerging Good Practices. Executive Summary, Project Background, Conclusions and Recommendations landed between these two groups. Notably, while Findings scored in the lower band, reviewers consistently noted that findings sections were clearly organized — the gap appears to lie in completeness (e.g. addressing unintended outcomes and more consistent citing of evidence) rather than presentation.

Key areas for improvement - As in prior years, the main shortcomings echo those identified in 2024. Within Evaluation Methods, the lowest-scoring items were sampling procedures, adherence to UNEG norms and ethical considerations, and the extent of participatory practices. ILO's continued emphasis on the latter appears warranted: the evaluations with the highest participatory scores also recorded the highest mean scores for overall quality. Within Project Background, reviewers most frequently flagged stakeholder mapping and the description of project context as needing to be more fully developed.

Persistent issues in other components included unintended outcomes not being addressed in Findings, stakeholder consultation not being evident in the development of Recommendations, and the Lessons Learned were often restatements of findings rather than transferable insights. Annex completeness was also an issue with reports often missing data collection instruments, interviewee lists, or document review lists.

In respect to cross-cutting themes, consistent with 2024, disability inclusion remains the cross-cutting theme most frequently absent from evaluation frameworks. However, improvement was seen elsewhere: consideration of SDGs was more frequently part of evaluation questions, and while environmental sustainability and just transition considerations remain highly variable across regions and themes, they were flagged far less often than disability inclusion. Coverage of pandemic-related themes, a concern in 2024, was not raised at all in the 2025 portfolio.

UN-SWAP performance - While half of all evaluations met UN-SWAP requirements and the other half approached them, with no evaluations missing requirements, performance was uneven across the three criteria. Gender considerations were most always embedded in the evaluation's scope and design but were not consistently carried through with the same rigor into methodology, findings and recommendations.



ilo.org

International Labour Organization
Route des Morillons 4
1211 Geneva 22
Switzerland