International Labour Organization

Evaluation Office

i-eval THINK Piece, No. 21

Results and reflections from a quality appraisal of ILO evaluations, 2020

By Universalia Management Group
November 2021

Results and reflections from a quality appraisal of ILO evaluations, 2020

Think Piece, No. 21

Universalia Management Group
ILO Evaluation Office

November 2021

Contents

- 1. INTRODUCTION 1
- 2. FINDINGS 2
- 3. CONCLUSIONS 7

1. Introduction

1.1 Background

As part of the ILO's Evaluation Office's (EVAL) longstanding quality assurance process[1] the Universalia Management Group Limited (hereinafter, "Universalia" or "the review team") conducted a Quality Appraisal (QA) of 46 independent development cooperation project evaluations submitted to the ILO's Evaluation Office between January 2020 and December 2020.

This Think Piece reports a summary of the 2020 findings, conclusions and recommendations. In addition to the introduction, the report contains two sections:

- Section 2 presents the main findings of the assessment, and
- Section 3 presents the main conclusions of the quality appraisal.

> In 2020, the ILO Evaluation Office (EVAL) shifted to a rolling quality assurance process of decentralized project evaluations. Universalia has been appraising these evaluations on a biweekly basis to allow EVAL to identify and act upon emerging quality issues rapidly. Results are disseminated throughout the ILO's evaluation network so that interventions can be made in a targeted manner.

1.2 Overview and implementation of the quality appraisal tool

The ILO's quality appraisal tool looks at four different dimensions structured in four sections, allowing the reviewers to collect quantitative data on the quality of ILO's evaluation reports.

[1] This is one is a series of QAs. A total of ten quality appraisals have been conducted over the years of the independent evaluation reports submitted to EVAL. The results have been described in various Think Pieces, most recently in Robertson and Schroter, "Leveraging appraisal findings to improve evaluation quality", *i*-eval Think Piece No.4, March 2014; Friedman and Blight, "External quality appraisal: Implications for evaluation quality and utilization", *i*-eval Think Piece No.8, December 2014; Watts, "Quality assessments of ILO project evaluations: What are the next steps to better evaluations?", *i*-eval Think Piece No. 10, March 2016 ; Llabres, "Evaluation quality assessment methodology in the UN system and changes to the ILO's quality appraisal methodology", *i*-eval Think Piece No. 12, December 2017; Bustamante, López and Román, "ILO evaluation quality: Challenges and potential strategies", *i*-eval Think Piece No. 13, June 2018; Gonzales and Pénicaud, "Quality assessments of ILO project evaluations: Sustaining recent improvements", *i*-eval Think Piece No. 17, December 2019; Franche, Gonzales and Pénicaud, "Quality assessments of ILO decentralized evaluations: Key results of the quality appraisal 2019 and way forward for the integration of gender equality and empowerment of women considerations into evaluation", *i*-eval Think Piece No. 19, December 2020.

First, the tool captures descriptive data on demographic variables of each evaluation report, such as the region, department and year. Collected data can consequently be analyzed through the aggregation and identification of trends across these independent variables.

Secondly, the QA tool requires the reviewers to rate the quality of the content of the evaluation reports according to 58 different items (or criteria) grouped across the 10 standard sections that should structure an evaluation report.

Third, the comprehensiveness section of the tool ensures that data is collected on the presence or absence of key components that must be included in the report using a two-point scale (absent-present).

Finally, the UN-SWAP assesses four different items, in alignment with the Guidance on Integrating Human Rights and Gender Equality in Evaluation.

The QA covered 46 decentralized independent evaluation reports produced worldwide in 2020. The 46 evaluations conducted in a decentralized manner by EVAL with the help of certified evaluation managers included project, thematic, sector and clustered evaluations.

The sample included 35 final evaluations and 11 mid-term evaluations of projects from seven departments and all regions as well as interregional evaluations. The main purpose of the QA annual report was to provide a cumulative analysis of the evaluations submitted in 2020 and assess trends and comparisons with previous quality appraisals. The Quality Appraisal Summary Report informed the ILO's latest Annual Evaluation Report for 2020-2021, which was released in October 2021.

The process was implemented by two reviewers that appraised every single evaluation report to ensure inter-observer consistency. Once all reports of a given reporting period were appraised, quantitative ratings and qualitative information justifying the rating were aggregated in an excel sheet and overall scores were calculated for the quality, comprehensiveness and UN-SWAP dimensions. Aggregated scores and individual ratings were analyzed using quantitative and qualitative methods. Finally, an online survey was disseminated to a sample of evaluation managers.

2. Findings

2.1 Overall quality score

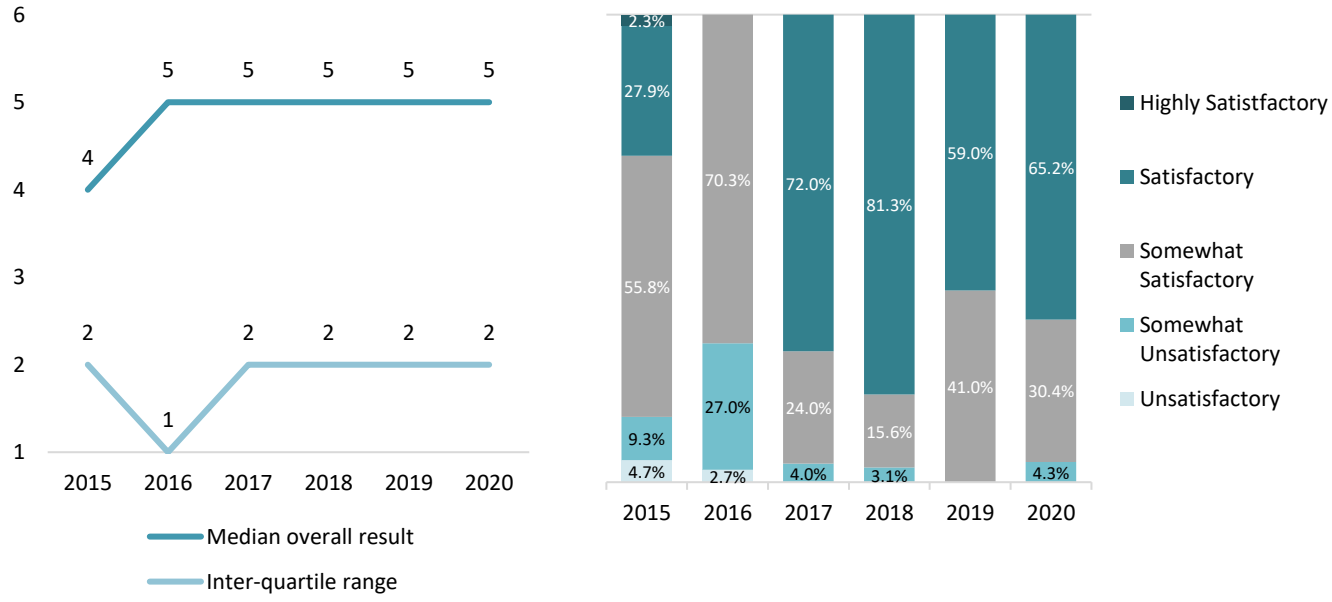
The quality of appraised reports has reached satisfactory ratings over the last six years. As illustrated in Figure 1, the median score for reports undertaken in a given year has remained at 5 since 2016.[2] The inter-quartile range, which measures the dispersion of results between evaluation reports for a given year, remained stable between 2015 and 2020, with 2016 being

[2] The overall scores are calculated by aggregating the ratings obtained for all items pertaining to the "quality" dimension of the QA, thus excluding the comprehensiveness and UN-SWAP dimensions. The results of the UN-SWAP assessment are presented separately.

the only year in which the dispersion appeared to be lower. Overall, the dispersion of ratings remained low, suggesting a certain homogeneity in the quality of reports over the years.

The proportion of reports receiving a "satisfactory" rating (5 out of 6) increased from 28% in 2015 to 65% in 2020. None of the reports received a "highly satisfactory" rating (6 out of 6) between 2016 and 2020, with 2015 being the only year for which some reports received this rating. Finally, while no report obtained a "somewhat unsatisfactory" rating (4 out of 6) in 2019, 4% of reports received this rating in 2020, reflecting likely the difficult circumstances under which these evaluations had to take place.[3] [4]

Figure 1. Overall ratings and evolution per year[5]



All evaluation reports submitted to EVAL are categorized by department, country and region. The analysis found that the quality is consistently high across the project evaluations of nine departments and offices represented in the sample, with five departments obtaining a median score equivalent to a "satisfactory" rating (5 out of 6) for the evaluations that were conducted in 2020.

The analysis revealed that the overall median score of evaluations is also consistent across regions. Evaluations have a median score considered "satisfactory" (5) in all regions except for Asia Pacific, where evaluations have a median score considered "somewhat satisfactory" (4).

[3] http://www.ilo.ch/wcmsp5/groups/public/---ed_mas/---eval/documents/publication/wcms_744068.pdf

[4] http://www.ilo.ch/wcmsp5/groups/public/---ed_mas/---eval/documents/publication/wcms_757541.pdf

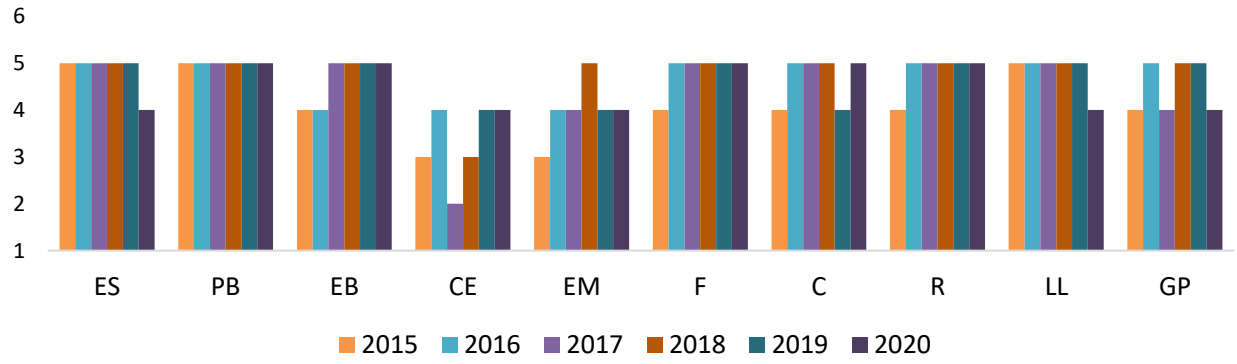
[5] Results for 2015 and 2016 are taken from the QA conducted in 2017 by Artival. Results for 2017 and 2018 are taken from the QA conducted in 2019 by Universalia. Results for 2019 include all the evaluation reports that were reviewed during the previous and the current QA by Universalia.

2.1 Quality of individual components

This section presents first an overview of the ratings obtained per component per year followed by an individual analysis for each individual component of the QA tool.

As illustrated in the figure below, five out of ten components obtained "satisfactory" ratings in 2020 (median rating of 5 out of 6), and five components ('Executive Summary', 'Criteria & Questions', 'Evaluation Methodology', 'Lessons Learned' and 'Good Practices') obtained a "somewhat satisfactory" median rating (4 out of 6). The 'Criteria & Questions' component remains the weakest component over the years, as observed in previous QAs. In 2020, quality decreased in three areas: 'Executive Summary', "Lessons Learned' and 'Good Practices' sections. Both obtained a median rating of 5 out of 6 in 2019 (satisfactory), and 4 out of 6 in 2020 (somewhat satisfactory).

Figure 2. Evolution of median rating per component and year



*Executive summary **(ES)**, Project background **(PB)**, Evaluation background **(EB)**, Criteria & questions **(CE)**, Evaluation methodology **(EM)**, Findings **(F)**, Conclusions **(C)**, Recommendations **(R)**, Lessons learned **(LL)**, Good practices **(GP)**.*

2.2. Comprehensiveness of evaluation reports

The QA tool includes a section on the comprehensiveness of evaluation reports whose purpose is to assess the overall structure of independent evaluations and to report any missing element. The reviewers thus utilized a binary scale to quantify the presence or absence of specific components expected to be found in an evaluation report.

Evaluation reports appraised generally include EVAL's mandatory components that are required in an evaluation report. As indicated in the table below, the ratings slightly increased over the years but are mostly stable (the minimum average was 87% in 2015, and the maximum was reached in 2019 with 92%).

QA results on comprehensiveness of evaluation reports (2020)

ITEMS	2015	2016	2017	2018	2019	2020
Title page using EVAL's template	67.4%	73.0%	92.0%	90.6%	94.9%	100.0%
Table of Contents	97.7%	94.6%	96.0%	100.0%	100.0%	100.0%
List of tables, figures and charts	32.6%	37.8%	36.0%	43.8%	53.8%	39.1%
List of annexes	79.1%	94.6%	92.0%	87.5%	94.9%	97.8%
List of acronyms or abbreviations	100.0%	91.9%	100.0%	93.8%	100.0%	97.8%
Executive Summary	100.0%	97.3%	100.0%	96.9%	97.4%	95.7%
Project Background	100.0%	97.3%	100.0%	100.0%	100.0%	100.0%
Evaluation Background	100.0%	100.0%	100.0%	100.0%	100.0%	97.8%
Criteria and Questions	72.1%	83.8%	72.0%	75.0%	82.1%	93.5%
Methodology	100.0%	97.3%	100.0%	100.0%	100.0%	100.0%
Findings	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Conclusions	95.3%	100.0%	92.0%	93.8%	100.0%	97.8%
Lessons learned	97.7%	100.0%	100.0%	100.0%	97.4%	97.8%
Emerging Good Practices	81.4%	89.2%	84.0%	84.4%	84.6%	84.8%
Recommendations	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Annexes	69.8%	59.5%	76.0%	65.6%	66.7%	58.7%
Average	**87.1%**	**88.5%**	**90.0%**	**89.5%**	**92.0%**	**91.3%**

2.3. SWAP

Universalia also analyzed the extent to which the ILO met UN-SWAP requirements, in 2020. The ILO is one of 69 organizations that are mandated to report against United Nations System-Wide Action Plan on Gender Equality and the Empowerment of Women (UN-SWAP-GEEW).

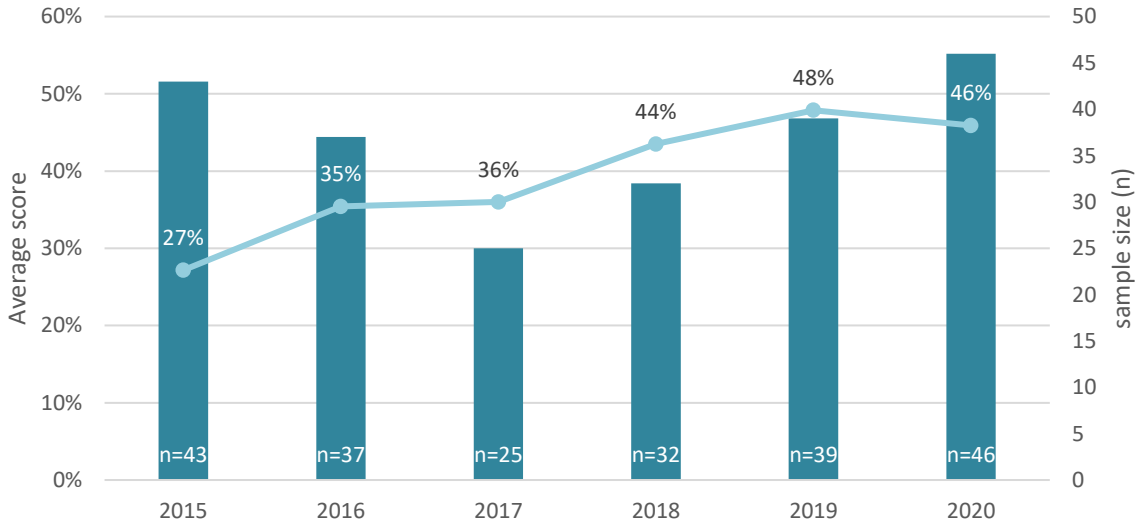
Reports are submitted on an annual basis using the UNEG endorsed Technical Note and related scorecard to report against the Evaluation Performance Indicator (EPI).[6] In order to comply with this requirement, on an annual basis, ILO asks quality assurance (QA) consultants to rate the EPI contained in the scorecard in compliance with the instructions

[6] UNEG. April 2018. UN-SWAP Evaluation Performance Indicator Technical Note.

found in the UN-SWAP-GEEW Technical Notes. The goal is for all UN system entities to "meet requirements" related to the EPI in terms of integrating GEEW in their respective evaluations. The 2020 reports assessed (46 in total) scored an average rating of 4.13 points. According to the criteria established in the UN-SWAP EPI (2018 version) and the aggregation of the scored obtained through the review process, ILO approached the UN-SWAP requirements in 2020.

Since the rating scale to calculate the meta-score went from being a 12-points scale to a 9-points scale with the 2018 revision of the Technical Note, Figure 3 shows the adjusted scores, presented in percentage of the maximum number of points that could be obtained every year since 2015.[7] The line clearly shows a positive trend in the extent to which ILO evaluations mainstream GEEW consideration in their reports, despite a very slight decrease between 2019 and 2020.

Figure 0. Adjusted meta-scores obtained between 2015 and 2020 (%)



The clear blue line clearly shows a positive trend in the extent to which ILO evaluations mainstream GEEW consideration in their reports. While Figure 3.1 can give the impression of a downward trend in the extent to which GEEW was mainstreamed in 2019 and 2020 evaluation reports compared to previous years, Figure 3.2 clearly shows an increase in the average meta-score obtained between 2015 and 2020.

[7] The meta-scores obtained between 2015 and 2018 were divided by 12 (the maximum possible meta-score based on the 2014 version of the scorecard) while the meta-score obtained in 2019 was divided by 9 (the maximum possible meta-score based on the 2018 version of the scorecard).

3. Conclusions

Since March 2020, the COVID-19 pandemic has affected the conduct of evaluations by ILO departments and offices. Several evaluation managers noted the impact of the global health crisis on the timeframe of evaluations, the methodological challenges that arose, and the difficulty of setting up virtual data collection missions to meet stakeholders and conduct participatory evaluation processes.

Nevertheless, all evaluation reports appraised in 2020 obtained ratings equal to or above "somewhat satisfactory", and there were no significant discrepancies in the overall quality of evaluation reports.

When assessing the quality of individual components, 'Project Background' is the only component which has received "satisfactory" ratings consistently since 2015. Items assessed under the 'Evaluation Background', 'Findings', 'Conclusions' and 'Recommendations' components were also positively appraised in most 2020 reports. The 'Criteria & Questions' component remains the weakest component over the years, as observed in previous QAs. In 2020, quality decreased in three areas: 'Executive Summary', 'Lessons Learned' and 'Good Practices' sections.

Generally, weaknesses in reports appear to indicate a lack of rigour on the part of evaluation teams to fully align with EVAL guidelines and checklists. Some sections could be further developed, in particular by including key methodological elements mentioned in inception reports. The linkages between evaluation questions, data collection methods, findings, conclusions, and recommendations could be made more explicit in many reports. Finally, one last area for improvement noted in reviewed evaluation reports is that lessons learned and good practices are often not developed as "stand-alone" documents that are easily understandable by an audience that did not read the full evaluation report. One way to facilitate the reading of lessons learned and good practices could be for evaluators to systematically include a clear and concise statement of the key message at the beginning of each lesson and practice.