i-eval THINK Piece, No. 17

Quality assessments of ILO project evaluations: Sustaining recent improvements

Juan-David Gonzales and Sophie Pénicaud
(Universalia Management Group)

International Labour Organization
Evaluation Office

December 2019

Quality assessments of ILO project evaluations:

Sustaining recent improvements

Juan-David Gonzales and Sophie Pénicaud
(Universalia Management Group)

December 2019

The responsibility for opinions expressed in this document rests solely with the author. The publication does not constitute an endorsement by the International Labour Organization.

This document has not been subject to professional editing.

International Labour Office
Evaluation Office

Table of Contents

List of acronyms

DEFP	Departmental Evaluation Focal Point
EM	Evaluation Manager
EVAL	ILO’s Evaluation Office
GEEW	Gender Equality and Empowerment of Women
IEE	Independent Evaluation of ILO’s Evaluation Function
ILO	International Labour Organization
IQA	Independent Quality Appraisal
MOPAN	Multilateral Organisation Performance Assessment Network
QA	Quality Appraisal
REO	Regional Evaluation Officer
TOR	Terms of Reference
UMG	Universalia Management Group
UN-SWAP	United Nations System-wide Action Plan
UNEG	United Nations Evaluation Group

Introduction

This Think Piece offers a reflection on the quality of the International Labour Organization's (ILO) decentralized evaluations appraised by Universalia in 2019 in the context of the Quality Appraisal (QA) exercises commissioned by ILO in previous years. This document succinctly presents and discusses ILO's Quality Assurance system and tools and addresses the broader evaluation process on which ILO relies to operationalize its decentralized evaluation system.

Following this introduction, the present document is organized around four main sections:

- **ILO's decentralized evaluation process**: The first section provides a rapid overview of ILO's independent evaluation function, of its hybrid decentralized evaluation process/system and presents the authors' perspective on this process within the context of the new Evaluation Policy.
- **ILO's Quality Appraisal system and tool**: The second section presents the QA approach and tool and describes changes to the tool introduced in 2019.
- **Key results of the Independent Quality Appraisal 2019**: The third section provides a summary of the latest quality appraisal (QA) report thus highlighting key improvements in evaluations and remaining weaknesses evaluators need to address when preparing their reports. It also discusses key factors that seem to have an influence on the quality of ILO's evaluation reports.
- **Emerging challenges and opportunities**: This final section identifies challenges and opportunities that the authors identified through implementation of the QA process in 2019.

QA Process 2019: Towards a rolling quality appraisal

ILO's Evaluation Office (EVAL) commissioned a total of eight batches of quality appraisals of the independent evaluation reports produced in the last few years. The most recent QA was conducted by Universalia between February and July 2019 and covered a sample of 64 final and midterm decentralized evaluation reports conducted worldwide between January 2017 and May 2019. The sample was representative of evaluations conducted by ILO's offices, including project, thematic, sector and cluster evaluations.

The main purpose of the QA was to provide a cumulative analysis of the evaluations submitted during the scope mentioned earlier and assess trends and comparisons with previous quality appraisals. The Quality Appraisal Summary Report informed ILO's latest Annual Evaluation Report for 2018-2019, which was released in September 2019.[1] This year's QA was also the opportunity for EVAL to launch, with the support of Universalia, a rolling QA system in which QA scores are provided for each individual report, soon after completion, thus allowing EVAL to detect quality issues quickly and to take immediate action.

The QA was conducted following a mixed-methods approach, combining quantitative analysis, using EVAL's scoring tool, and qualitative analysis to identify common trends, strengths and weaknesses observed within the sample of reports. Data collection methods included a thorough review of the quality appraisal system and EVAL's tool, a four-step appraisal process, and an online survey disseminated to a sample of evaluation managers.

[1] EVAL. October 2019. *Annual Evaluation Report 2018-2019*.

1 ILO’s decentralized evaluation system and process

ILO’s independent evaluation function

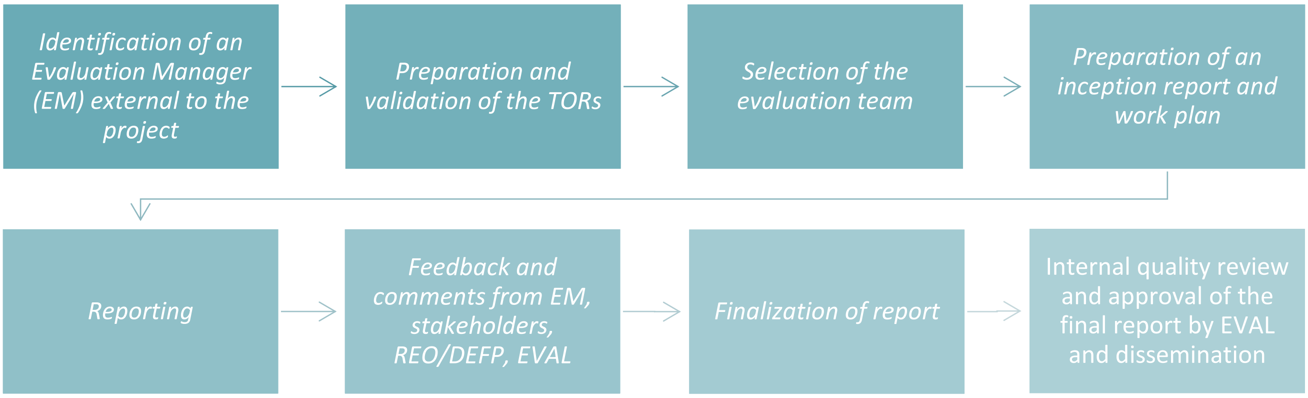
ILO’s evaluation function operates as a completely separate office of evaluation (EVAL), reporting directly to the Director-General’s office, thus guaranteeing the independence of the function. EVAL is the central body responsible for independent evaluations of ILO’s strategies, policies and programmes. The office is supported by Regional Evaluation Officers (REOs) in five regional offices, by Departmental Evaluation Focal Points (DEFPs) and by evaluation managers (EMs). REOs have dual reporting lines. Administratively, they report to regional management, and technically, to EVAL. This structure lays the foundation of EVAL’s hybrid decentralized system

A rapid overview of ILO’s hybrid decentralized evaluation system and process

A vast majority of evaluations appraised over the years have been independent project evaluations and decentralized evaluations. According to ILO’s evaluation policy (2017), independent evaluations and reviews are managed by EVAL or independent ILO officials, overseen by EVAL and carried out by EVAL officers or external independent evaluators. Decentralized evaluations include thematic evaluations (other than those managed by EVAL), project evaluations, impact evaluations, joint evaluations and internal reviews, which also include self-evaluations.[2]

ILO’s hybrid decentralized system follows the 8-step process presented in Figure 1 below to implement the above-mentioned evaluations.[3]

Figure 1 Implementation Process for Independent Evaluations


Identification of an Evaluation Manager (EM) external to the project
Preparation and validation of the TORs
Selection of the evaluation team
Preparation of an inception report and work plan
Reporting
Feedback and comments from EM, stakeholders, REO/DEFP, EVAL
Finalization of report
Internal quality review and approval of the final report by EVAL and dissemination

Given the overall good quality of evaluations appraised by Universalia (further discussed in section three), the reviewers consider that the process currently utilized by ILO is fit for purpose. Further evidence to support this idea can be found in the *Analysis of the Evaluation Function in the United Nations System* conducted by the Joint Inspection Unit in 2014 that observed that ILO had a well-institutionalized evaluation function.[4] This is also supported by the results of the 2015-16 Multilateral Organisation Performance Assessment Network (MOPAN)

[2] ILO. 2017. *Evaluation Policy*, pg. 38-39.

[3] Based on the ILO’s policy guidelines for evaluation (2017) and on Universalia’s own experience working with ILO EVAL.

[4] JIU. 2014. *Analysis of the Evaluation Function in the United Nations System*. Geneva.

Assessment that rated the ILO's application of evidence-based planning and programming as being satisfactory.[5] The *2016 Independent Evaluation of ILO's Evaluation Function* (IEE) furthermore observed that ILO had established "highly structured systems and processes to deliver the evaluation function" allowing for compliance with UNEG norms and standards.[6] These findings are laudable given the fact that ILO is considered as a medium-size organization (in terms of annual budget), and that the budgets for the appraised evaluations seemed to be generally low.

A perspective on ILO's evaluation process in the context of the new Evaluation Policy

Since the publication of the above-mentioned assessments and evaluations, a new Evaluation Policy (2017) was adopted by the ILO Governing Body in order to continue strengthening the accountability and learning purpose of its evaluation function. The Evaluation Strategy 2018-21 (2018) was adopted subsequently to operationalize the new 2017 Evaluation Policy. The new strategy notably aspired to align the evaluation function with the revised UNEG Norms and Standards for Evaluation (2016) and the findings of the 2016 IEE. The latter required EVAL to enhance evaluation methods to cover more adequately the specific mandate of ILO (social dialogue, tripartism, normative work). Some of the more specific objectives of the strategy were also to improve the quality of decentralized evaluations and to transition towards the use of strategic cluster evaluations.

To our knowledge, the current structure, policy and strategy remain appropriate to ensure compliance to the general and institutional norms laid out in the revised UNEG Norms and Standards document. Taking a more specific look at the UNEG Norms and Standards, we note that the process described above, a process anchored in an independent evaluation office and whose evaluation products are quality assured by an external and independent entity, ensures compliance with all five overarching standards.[7]

Management of evaluation function (standard 2) has, for example, been further strengthened by issuing updates of the *ILO Policy Guidelines for Evaluation* (edition 3, 2017) that are intended to provide a complete package of guidance for ILO staff involved in planning, managing or overseeing evaluations. Updated guidance notes, checklists and templates are also made available to evaluators and evaluation managers. The updated guidance notes seem to be useful and well-designed despite the fact that some notes are still outdated. They are useful tools that find no equivalent, to our knowledge, in other evaluation units in the United Nations.

In terms of competencies and ethics (standard 3), while the QA exercise did not form an objective judgment on the competencies of managers and consultants, we concur with the finding of the IEE that it is difficult to find evaluators with experience in required technical areas of normative work and tripartite working structures, and knowledge of specific geographical regions, along with the required linguistic skills.

Regarding the conduct of evaluations (standard 4), ILO made mandatory the integration of International Labour Standards and gender equality in all evaluations (standard 4.7 on human rights-based approach and gender mainstreaming strategy). While Universalia's QA did note an improved integration of these cross-cutting policy drivers in the scope of evaluations (as an evaluation criterion or question) there is no clear or explicit explanation of the methodologies used to integrate a broader human rights-based approach beyond core labour rights in evaluations or in ILO tools.[8] Also, we did not notice, in the TORs or in evaluation reports, a clear or explicit emphasis on evaluability assessments (standard 4.2 on evaluability assessment).

[5] MOPAN. 2017. *MOPAN 2015-16 Assessments – International Labour Organization Institutional Assessment Report*.

[6] UNEG. 2016. *Norms and Standards for Evaluation*. New York: UNEG.

[7] Institutional framework, management of evaluation function, evaluation competencies, conduct of evaluations, quality. UNEG. 2016. *Norms and Standards for Evaluation*. New York: UNEG.

[8] Other UN Agencies (e.g. OHCHR, UNICEF, UNFPA, UNDP) place emphasis on a diverse set of HR concepts related to participation, inclusion, equity, accountability, confidentiality, disaggregation in their evaluations.

The existence of an embedded quality control process in the evaluation process presented above, and of an external ex-post quality assurance system, guarantees compliance to UNEG quality standard (standard 5).

Regarding the institutional framework (standard 1) that encompasses a number of items related to the general norms for evaluation (independence, credibility, usefulness), accompanied by the appropriateness of resource allocation to the evaluative function, we believe some limitations presented in previous assessments of ILO’s evaluation function remain valid.

2 ILO’s Quality Appraisal system and tool

A rapid overview of the tool

A previous *i*-eval Think Piece prepared by I. Llabres (2017) described the changes made by ILO to its evaluation quality appraisal tool and system in recent years that led to the adoption of the current system. In its current format, ILO’s quality appraisal tool looks at four different dimensions structured in four sections, namely:

1) **Demographics**: collects descriptive data on variables of each evaluation report, such as the region, department and year;
2) **Quality**: collects data based on the quality rating given by the reviewer on the different components and items that must be included in each evaluation report based on a six-point scale (from highly unsatisfactory to highly satisfactory);
3) **Comprehensiveness**: collects data on the presence or absence of key components that must be included in the report using a two-point scale (absent-present);
4) **UN System-wide Action Plan (UN-SWAP) on Gender Equality and the Empowerment of Women** collects data on the extent to which Gender Equality and Empowerment of Women (GEEW) has been included in ILO’s evaluations using a four-point scale.

The four dimensions of the quality appraisal tool are designed to allow the reviewers to collect quantitative data on the quality of ILO’s evaluation reports. Collected data can consequently be analyzed through aggregation and identification of trends and extreme values using different independent variables (years, departments, regions, etc.).

The three more substantial dimensions of the QA system are covered in the *quality*, in the *comprehensiveness* and in the *UN-SWAP* sections of the tool. More specifically, the *quality* section 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. The *comprehensiveness* section determines the thoroughness of the structure of the report itself: it collects data on the presence or absence of key components that must be included in the report using a two-point scale (absent-present). The *UN-SWAP* on the other hand assesses four different items, in alignment with the Guidance on Integrating Human Rights and Gender Equality in Evaluation (2014).[9]

QA implementation process

While there is no standard process to conducting Quality Assurance work, the review team followed a four-step process in which each evaluation report was appraised by two reviewers to ensure the reliability of findings and inter-observer consistency. First, one reviewer conducted an in-depth appraisal of an evaluation report, and attributed ratings, provided justifications, and offered other relevant comments as needed. Then, a second reviewer validated or challenged the assessment made by the first reviewer by raising questions or additional arguments regarding different ratings. The final decision on the ratings was taken by the first reviewer.

Finally, qualitative data and quantitative scores were aggregated in an excel sheet. Once compiled, the reviewers identified any inconsistencies (where comparable observations for specific criteria had been associated with different scores or where the same scores were associated to different observations), reassessed each instance, and adjusted the score if required.

[9] UNEG. 2014. *Guidance on Integrating Human Rights and Gender Equality in Evaluation*.

Changes introduced in 2019

Universalia proposed and introduced a few modifications to the QA tool in order to improve the quality and reliability of the data generated through the QA process.

First, the reviewers emphasized the importance to assess the inclusion of each of ILO's cross-cutting policy drivers separately. As such, instead of having one single item to evaluate cross-cutting policy factors all together, Universalia reviewers modified the tool by adding specific items to assess the inclusion of '*Tripartism and Social Dialogue*', '*International Labour Standards*' and '*Environmental Sustainability*'. However, the review team recommends to future reviewers to always ensure that the most recent cross-cutting policy drivers (presented in the Director-General's Programme and Budget) are included in the tool.

Second, Universalia as requested updated the UN-SWAP dimension by ensuring its alignment with the most recent UN-SWAP Evaluation Performance Indicator Technical Note (2018) published by the United Nations Evaluation Group (UNEG). The main implication of this change was that individual evaluation reports would be appraised according to three rather than four scoring criteria.[10] Again, the review team recommends that each reviewer ensure conformity of the ILO QA tool with the most recent version of the UN-SWAP scoring criteria and its corresponding definitions.

Third, Universalia reviewers have proposed and began testing the inclusion of qualitative evidence when appraising individual evaluation reports. As such, the review team is now providing a short summary of the key strengths and weaknesses of each appraised evaluation report as a way to provide an explanation, beyond the quantitative ratings, of the quality of each report. It thus becomes possible for external readers to rapidly understand, for each evaluation report, why a given score has been attributed to an evaluation. We believe that including a qualitative assessment can complement numerical data through interpretative feedback and explanation. Such feedback can be useful to EVAL, evaluation managers, and evaluators upon which ILO regularly relies to conduct decentralized programme evaluations.

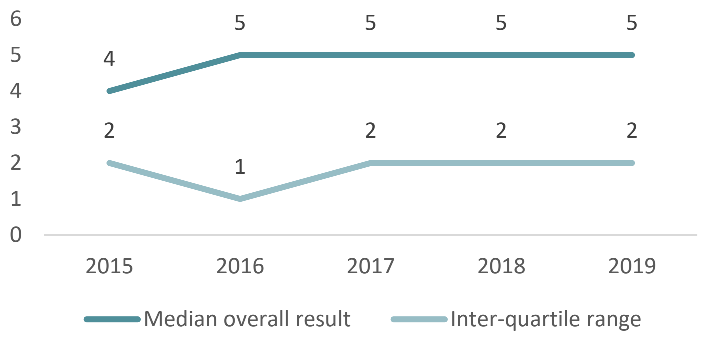
Depending on the needs of ILO, transitioning to a QA model that would require reviewers to provide more systematically qualitative comments along with their quantitative ratings (for each dimension or for a cluster of elements) could serve to provide more robust and precise feedback on the strengths and weaknesses of evaluations.

[10] UNEG. April 2018. *Guidance Document: UN-SWAP Performance Indicator Technical Note*.

3 Key results of the Quality Appraisal 2019

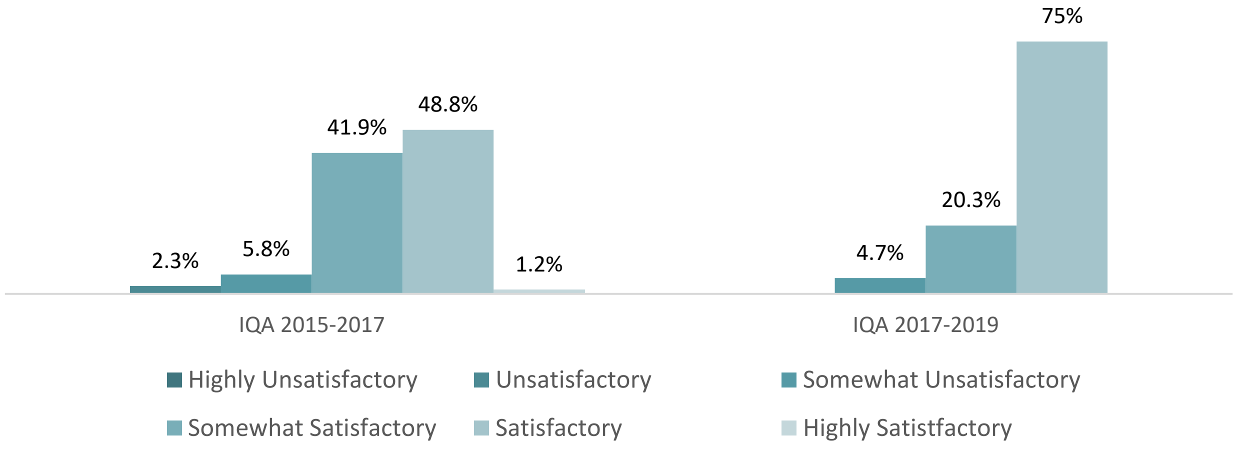
The quality of appraised reports has reached satisfactory levels over the last four years. As illustrated in Figure 2, the median score for reports undertaken in a given year remained satisfactory since 2016.[11] The inter-quartile range, which measures the dispersion of results between evaluation reports for a given year, remained stable between 2015 and 2019, 2016 being the only year in which the dispersion appeared to be lower. Overall, the dispersion of scores remained low, suggesting a certain homogeneity in the quality of reports over the years.

Figure 2 Overall ratings and evolution per year



The results of the last QA demonstrated a significant increase in the proportion of reports that obtained ratings equal or above somewhat satisfactory ratings. As demonstrated in the Figure 3 below, while 91.9% of reports pertained to that category in 2015-2017, nearly 95.3% obtained these ratings in 2017-2019, representing a 3.4% increase. However, no report has obtained a highly satisfactory rating during the last QA exercise, indicating there is still room for improvement.

Figure 3 Overall ratings and evolution per IQA



There were no significant discrepancies in median scores obtained when aggregating results per regions: there were no discrepancies between ILO's ten departments and between regional offices that commissioned evaluation reports in 2017-2019. Evaluation reports obtained a median score equivalent to a satisfactory rating regardless of

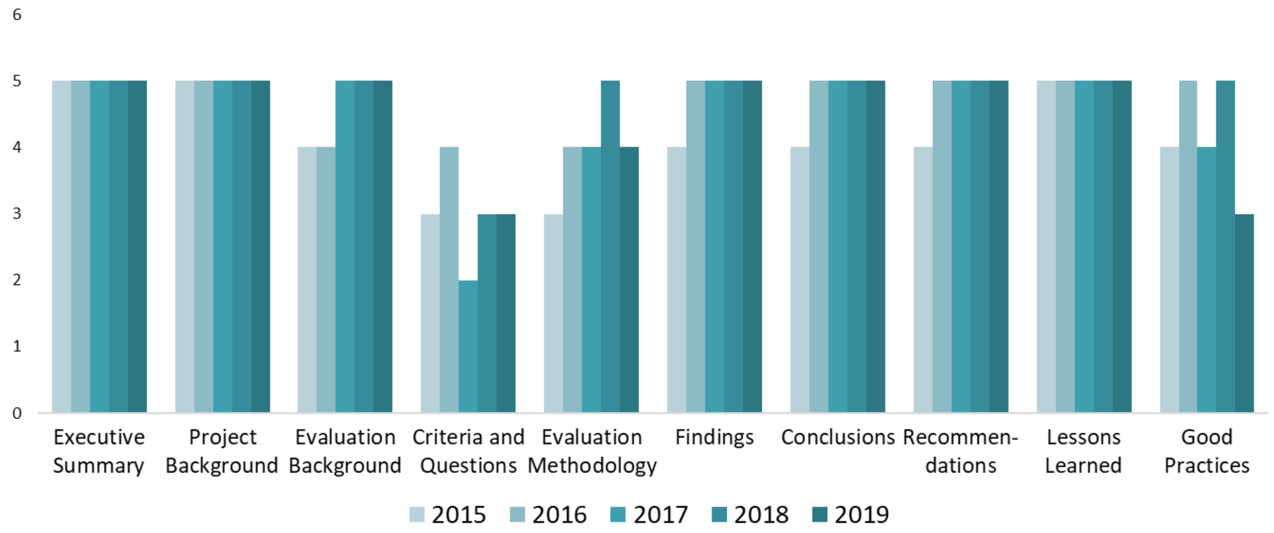
[11] The overall scores are calculated by aggregating the scores obtained for all items pertaining to the "quality" dimension of the QA, thus excluding the comprehensiveness and UN-SWAP dimensions. The results of the UN-SWAP assessment are presented separately.

where the evaluations were conducted, and regardless of how many evaluations were conducted in a specific region. However, inter-regional differences were observed when comparing median results obtained for specific sections of reports. The results of the QA, for example, showed that evaluation reports conducted for Africa and Europe regional offices were weaker in terms of providing strong and complete evaluation matrices and identifying good practices.

As illustrated in Figure 4 below, a yearly comparison shows that the strongest sections in evaluation reports have been the executive summary, the project background description and the lessons learned identified by evaluation teams. Findings sections were also positively reviewed and addressed all evaluation criteria in most reports. Another strength is the fact that most evaluations included their recommendations, lessons learned and good practices as per ILO EVAL's templates which is an improvement since 2015.

By contrast, the only section not reaching satisfactory requirements since 2015 is the section on evaluation criteria and questions. Evaluation matrices are an essential tool to ensure the thoroughness and quality of evaluation reports but were found to be missing in most evaluation reports appraised (2017-2019). In some cases, questions were found to be exactly the same as in the terms of reference, with apparently no efforts made in rephrasing, reorganizing, and differentiating key questions and sub-questions, including indicators and sources of data collection.

Figure 4 Evolution of median score per component and year

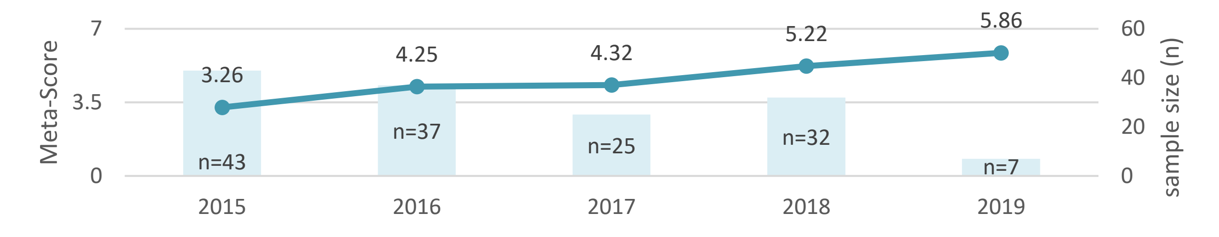


The evaluation methodologies adopted by independent evaluators or, more specifically, the way methodologies are presented in evaluation reports, were found to be one of the most significant weaknesses in the reports. Clear presentation and justifications of why specific evaluation approaches are selected were often missing or lacked specificity and remained generic for most evaluations. Explanations of sampling procedures and the rationale for the selection of stakeholders were also poorly developed in reports. This affected the overall robustness of data collection methods as fair participation of stakeholders was not clearly demonstrated, and little was said about the representativeness or rationale that led to the selection of interviewed stakeholders or of the regions that were visited. In addition, cross-cutting policy drivers (tripartism and social dialogue; international labour standards; environmental sustainability) were not sufficiently covered in most evaluation methodologies, a result that could be explained by the fact that EMs do not adequately insist on these aspects or do not share ILO's checklist to guide evaluators. Gender equality is, however, generally better integrated than other cross-cutting policy drivers in evaluations as it is often included as a mandatory evaluation question or as a specific evaluation criterion. That said, a more thorough description of gender considerations in the presentation of the object being evaluated, as well as methodologies that would be better suited to capture and analyze gender inequalities and inequities, would be desirable.

Other specific weaknesses included: the omission of details on evaluation funding arrangements and timeframes on the cover page of reports; the list of conclusions in executive summaries, often combined with the list of findings; vague targeted users in lessons learned and good practices sections; and unintended outcomes that were not mentioned in most reports.

Gender has been increasingly considered in evaluation reports. EVAL made significant efforts to incorporate gender into evaluations. Yearly comparisons of the average UN-SWAP Evaluation Performance Indicators (EPI) meta-scores from 2015 to 2019 demonstrate a slight but steady improvement over time (see Figure 4 below). Gender is better integrated into evaluation questions and appears to be more discussed under several findings of evaluations. However, gender should be better integrated into the evaluation scope of analysis, and evaluations should be more driven by gender-responsive methodologies to a greater extent.

Figure 5 Average meta-scores obtained between 2015 and 2019



Observations and lessons

The main factor that could explain the overall progression in the quality of evaluation reports since the last QA report (2015-2017) seems to be the better knowledge and understanding of guidelines and EVAL's requirements among evaluation managers[12]. However, the data obtained through the QA scoring tool and the online survey did not allow isolating any specific factors that could explain variations in the quality of reports produced. The reviewers also excluded the possibility that the differences in the scores obtained across years could be attributed to the fact EVAL changed QA provider/consultants.[13]

There is, for example, no clear correlation between the total ratings of the evaluation report and the budget used for the evaluations. In other words, evaluation budgets do not explain the variance in the evaluations' total rating. The results of the analysis of the relationship between evaluation timeframes and the total rating of the reports is also inconclusive even though the previous QA showed a statistically significant relationship between the time allocated to conduct an evaluation. The average timeframe of evaluations conducted between 2017 and 2019 was slightly longer than the average timeframe of evaluations conducted in 2015-2016, which could partially explain why the QA conducted in 2017 found a significant statistical correlation between evaluation timeframe and final rating, while the current QA did not. However, the review team believes tight budgets and timeframes are often perceived by external evaluators as key determinants of the overall quality of evaluations. This is further discussed in the last section of this think piece.

[12] The role of the internal quality control was not taken into account and may be something to assess in the future. Linking performance appraisal of SEOs to the ex-post quality of the evaluations they approved may have lead to a more stringent internal review

[13] A chi-squared test was conducted for the UN-SWAP dimension, rejecting the null hypothesis that the medians remained the same across time for two UN-SWAP criteria.

ILO is "becoming a more consistent and visible champion for gender mainstreaming and introducing good practices that are in some ways ahead of its UN evaluation peers".[14]

TheQA however observed a statistically significant relationship between an evaluation's score on the criteria and questions section, and the overall score of the evaluation reports for 2017-2019. Evaluation reports that included an evaluation matrix presented a higher average final score than reports without a matrix. In sum, the review team observed that the use of an evaluation matrix seems to be a determinant of the overall quality of the evaluation report, and evaluation reports that included an evaluation matrix received a higher final score on average. These results illustrate the importance of defining the evaluation criteria and questions and using an evaluation matrix as a framework for conducting evaluations.

However, including an evaluation matrix may not be a necessary condition to ensuring the quality of an evaluation report: findings can be appropriately developed, aligned to the proposed methodology, and be evidence-based, even if a report does not include a proper evaluation matrix. The reviewers do believe, however that well designed evaluation matrices significantly strengthen the overall quality of an evaluation report by facilitating the triangulation of data and strengthening the evidence base of findings. The design and integration of a structured evaluation matrix in each independent report should be mandatory and systematized. It should specifically include key questions rephrased from the ToR and additional sub-questions under each criterion with corresponding indicators, data collection methods and sources.

[14] ILO. December 2016. *Independent Evaluation of ILO's Evaluation Function 2011-2016*.

4 Emerging challenges and opportunities

ILO's hybrid decentralized evaluation process and system

Based on the observations made in section 1 of this report, as well as on the experience acquired in conducting the quality appraisal of ILO's evaluation in 2019, we identified the following challenges and opportunities.

- In terms of **independence**, as highlighted by IEE, the fact that Regional Evaluation Officers (REO) report to the regional management structure rather than directly to EVAL can still be perceived as conflicting with the independence of the evaluation process. Also, the fact that the EMs often know one another in the organization (even in larger regions) and only report temporarily to EVAL for the assigned evaluation, remains an issue in terms of perceived independence.
- In terms of **credibility**, the results of the current QA continue to highlight the general weakness of the methodological approaches used in the vast majority of evaluations. As indicated in the IEE, "the dominant evaluation approaches in ILO are somewhat conservative and focused on examining the achievement of results frameworks rather than on examining the underlying theories of change".[15] As such, the generic nature of methodologies employed by evaluators does not seem to fit with ILO's objective to enhance evaluation methods to assess the specific mandate (social dialogue, tripartism, normative work) of ILO. However, recent methodological guidance issued in September 2019 might reverse this trend.[16]
- **Usefulness** of evaluations can still be improved despite the significant progress achieved in recent years in the dissemination of evaluation results (*i*-track, *i*-eval Discovery). Increased utilization of strategic cluster evaluations was included in the new evaluation strategy to improve the utility of the evaluations. Among the 64 evaluations appraised by Universalia, eight seemed to be clustered around specific elements, for example: global programmes (Better Work), global outcomes (P&B Outcome 17), common themes and regions (sustainable enterprises in the Americas). While this is a good start, conducting clustered evaluations more frequently could be an effective and efficient way to improve utility of evaluations.
- In terms of **resources**, we believe that budgets allocated to evaluations remain low. This is a perception shared by international evaluators and that was highlighted by the IEE. The IEE noted that, while one would expect expenditure in the range of 1.5–2 per cent of the total organizational budget being allocated to evaluation, approximately half of that amount was being committed.
- The IEE concluded for example that the current structure that relies on the system of volunteer Evaluation Managers is a result of the shortage of M&E Specialists in the organization which, in turn, may also be related to the fact that the evaluation function was qualified as "structurally underfunded".[17]

In addition, the reviewers note that several of ILO's guidance notes that were designed to help EMs and evaluators improve the overall consistency and quality of the reports should be updated more frequently. Most of these guidance notes date back to 2014 and should be more closely aligned with most recent changes introduced in the new evaluation policy (2017) and strategy (2018). These guidance notes should also be closely aligned with the QA tool.[18]

In terms of clustering, we believe that conducting larger-scale evaluations could have multiple advantages, including:

- Drawing increased interest of stakeholders (including direct and indirect clients of the evaluation) to evaluations thus increasing the importance given to accountability and utilization of evaluation findings;

[15] ILO. December 2016. *Independent Evaluation of ILO's Evaluation Function 2011-2016*.

[16] ILO. September 2019. *Guidance Note: Adapting Evaluation Methods to the ILO's Normative and Tripartite Mandate*.

[17] ILO. December 2016. *Independent Evaluation of ILO's Evaluation Function 2011-2016*.

[18] EVAL is launching all revised guidance notes in early 2020.

- Clustering evaluations could stimulate knowledge sharing and increase communication across projects;
- Having larger budgets and longer timelines to implement more complex evaluations would attract more experienced evaluators that have the capacity to implement higher quality evaluations both in terms of methodology and usefulness;
- Decreasing the number of small evaluations can reduce the transaction costs related to their management, as well as reduce the risk of burdening stakeholders (such as tripartite constituents) that are commonly consulted in the context of smaller project evaluations.
- At this stage, we believe that the tool to conduct the QA of single evaluations is suited to appraising the quality of clustered evaluations.

ILO's Quality Appraisal system and tool

ILO made significant progress in recent years on moving towards the implementation of a rolling quality appraisal system. However, since the QA process is not embedded in the evaluation process or, in other words, since QA follows the internal approval of the evaluation reports, EVAL still needs to determine what actions to take when lower quality reports are identified.

A key opportunity to test in the following months will be for EVAL to identify recurrent issues and weaknesses and to determine whether they can act upon them more rapidly and in a more targeted fashion. EVAL is currently engaged in the preparation of a curriculum for an advanced EMCP training that could for example be tailored to address some of the most recent emerging issues related to the quality of evaluation reports.

Other relevant opportunities and recommendations were also formulated in previous Think Pieces commissioned by ILO and could deserve some attention to improve the QA system or tool (e.g. weighting QA dimensions or items in accordance with their relative importance).

Quality of evaluation reports

During the QA exercise, the review team observed several good practices which strengthen quality and thoroughness of reports. However, it appears that most of these practices are not systematically monitored by evaluation managers as they are not clearly explained or do not appear at all in EVAL guidance notes and checklists. As such, additional efforts should be made by ILO to ensure evaluators design sound evaluation methodologies from the inception phase in order to ensure that evaluation reports reach the highest quality standards, particularly on the following items:

- Independent evaluators should go beyond simply referencing a mixed-methods approach but should also propose clear and sound mixed methodological tools that are suitable to the purpose and objectives of the evaluation conducted.
- In terms of methodologies to better evaluate mandates that are specific to ILO, methodological innovation should be incentivized by allocating more resources to evaluators who engage in evaluative work by specifying, in the Terms of Reference, the need for them to propose innovative methodologies. An alternative approach for ILO could be to directly invest in the development of innovative methodological approaches, directly linked to ILO's mandate, that can become an institutionalized process within the organization. The recent guidance note on *Adapting Evaluation Methods to the ILO's Normative and Tripartite Mandate* is heading in this direction by encouraging the systematic integration of social dialogue and normative context in ILO's monitoring and evaluation products.[19]

[19] ILO. September 2019. *Guidance Note: Adapting Evaluation Methods to the ILO's Normative and Tripartite Mandate*.

- Similarly, a strong rationale should be provided to justify the use of specific data collection methods that correspond to distinct evaluation goals. In doing so, this strengthens methodological reasoning and subsequent implementation of the evaluation methods.
- The selection of provinces and populations under review, as well as the choice of respondents (by interview, focus group discussion, survey, etc.), should be systematically clarified and justified through clear sampling procedures, and a rationale for selecting stakeholders. These sections and their level of detail are essential to ensure the participation of the right set of relevant stakeholders in the evaluation process, including project beneficiaries, and more specifically women and vulnerable populations.
- Tripartism and social dialogue, International Labour Standards, environmental sustainability and gender considerations should all be treated as cross-cutting issues that are integrated throughout the report, rather than being covered in a separate section. During the inception phase, these issues should be systematically included in the questions, sub-questions and/or indicators under most evaluation criteria to ensure their proper integration.
- Theory of change, results framework or Intervention Logic of the assessed project should be presented in detail in the evaluation report and made available to evaluators during the inception phase of the evaluation process.