i-eval THINK Piece, No. 12

Evaluation quality assessment methodology in the UN system and changes to the ILO's quality appraisal methodology

Iker Llabres

International Labour Office
Evaluation Office

December 2017

Evaluation quality assessment methodology in the UN system and changes to the ILO's quality appraisal methodology

By Iker Llabres

December 2017

International Labour Office
Evaluation Office

Table of Contents

List of Acronyms

EVAL	Evaluation Office of the ILO
FAO	Food and Agriculture Organisation of the United Nations
IEE	Independent Evaluation of the ILO's Evaluation Function
ILO	International Labour Organization
JIU	Joint Inspection Unit
NA	Not applicable or missing components
QA	Quality Appraisal / Assessment
ToR	Terms of Reference
UN-SWAP	United Nations System-wide Action Plan on Gender Equality and the Empowerment of Women
UNDP	United Nations Development Programme
UNEG	United Nations Evaluation Group
UNEP	United Nations Environment Programme
UNESCO	United Nations Educational, Scientific and Cultural Organization
UNFPA	United Nations Population Fund
UNICEF	United Nations Children's Fund
UNODC	United Nations Office on Drugs and Crime
WFP	World Food Programme
WIPO	World Intellectual Property Organization

1. Introduction

The ILO places strong emphasis on ensuring that credible independent evaluations of its strategies, programmes, and projects are conducted in accordance with the expectations of its constituents and donors and that they are in compliance with international norms and standards (e.g. UNEG Standard 5.1). Apart from high-level evaluations undertaken at the strategy level and reported to the ILO's Governing Body, most of the Office's evaluation work exists at the project level and is integrated into a larger effort to ensure effective design, implementation, monitoring and reporting of projects.

A recent independent evaluation of the ILO's evaluation function (IEE) noted that the quality assessment undertaken in the portfolio analysis generally complied with UNEG quality standards. It went on to report that the quality assurance system had technical strengths, but also offered several opportunities for improvement.

This comment prompted the ILO's Evaluation Office (EVAL) to undertake a revision of its quality appraisal system. To help inform the revision, EVAL conducted a review of the quality appraisal systems of a sample of similar UN agencies, programmes, and funds (Section 2). In addition, EVAL conducted an analysis of the results of past quality appraisal exercises (Section 3). Based on these two sources of information, EVAL made changes to its quality appraisal system (Section 4). The current status of the ILO's quality appraisal system is described in Section 5.

1.1 How did EVAL do quality assessments in the past?

EVAL is the office responsible for implementing the ILO's evaluation policy. The policy identifies quality control as one of EVAL's roles (2017, p. 10). At the decentralized level, such a role involves assessing the quality of independent project evaluations. EVAL has a real time internal quality control system (provide details—various layers of controllers, manuals, and training) and ex-post external quality appraisals. The outcome of the quality appraisal (QA) is reported in the Annual Evaluation Report (AER) that the Director of EVAL presents to the Governing Body (GB). Counting the 2017 appraisal, a total of eight ex-post quality appraisals of the independent evaluation reports submitted to EVAL have been conducted. THINK Pieces have been published by EVAL summarising the findings of the three most recent appraisals (Friedman and Blight, 2014[1]; Robertson and Schroter[2], 2014; Watts, 2016[3]).

The appraisals are conducted by independent consultants who follow a methodology developed by EVAL. Depending on the number of reports, the entire population or a stratified (by language) random sample is appraised. The consultant(s) applied an 'appraisal tool' with 137 items dealing with elements of the report such as executive summary, background, findings, and recommendations. The tool was divided in two parts: 1) a **component** check that marks the elements of the reports as present or absent, and 2) a part that requires reviewers to rate the **quality** of an aspect of the element.

[1] Friedman, J. and Blight, N. (2014) External Quality Appraisal: Implications for evaluation quality and utilization. i-eval THINK Piece No. 8. Available at: http://www.ilo.org/eval/newsletter-and-think-pieces/WCMS_329163/lang--en/index.htm

[2] Robertson, K. and Schroter, D. (2014) Leveraging appraisal findings to improve evaluation quality. i-eval THINK Piece No. 4. Available at: http://www.ilo.org/eval/newsletter-and-think-pieces/WCMS_237914/lang--en/index.htm

[3] Watts, B. (2016) Quality assessments of ILO project evaluations: What are the next steps to better evaluations? i-eval THINK Piece No. 10. Available at: http://www.ilo.org/eval/newsletter-and-think-pieces/WCMS_329163/lang--en/index.htm

Each question or item was rated on a 4-point ordinal scale, where:

0 = unacceptable level of quality
1 = insufficient level of quality
2 = acceptable level of quality
3 = high quality.
NA

Finally, an index score was calculated for each section of the individual report and aggregated for the entire sample. A trend that was of concern to EVAL was that the average quality of reports remained stable at an 'acceptable' level over the last three quality appraisals despite initiatives to improve quality such as the Evaluation Manager Certification Programme, updated Evaluation Policy Guidelines, increased budgets to hire better quality consultants, etc.

2. Quality appraisals in the UN: A comparative analysis

To help inform the planned revisions, the ILO conducted a review of the quality appraisal systems of a sample of 11 similar UN agencies, programmes and funds. The organisations that were ultimately included in the sample were: FAO, UN Women, UNDP, UNEP, UNESCO, UNFPA, UNICEF, UNIDO, UNODC, WFP, and WIPO. The results are presented below:

The QA methodologies of the UN entities in the sample are compared in Table 1. The analysis focused on the number of items, their scale measurement, response categories, weighting and techniques for analysis. The results are summarized in Table 1. What is most interesting about the results is that quality appraisal systems in the United Nations are not harmonized.

Table 1. Quality Appraisal in the UN System

UN Entity	Scale of measurement	Response categories	NA score	No. of items	No. of sections	Weighted sections	Measure of central tendency
FAO	Ordinal[4]	6	Yes	28	9	Not mentioned	Not mentioned
ILO	Ordinal	6	Yes	58	10	No	Modes, medians
UN Women	Ordinal	4	No	39	8	Intrinsically	Qualitative
UNDP	Ordinal	6	No	55	6	Yes	Qualitative, customized
UNEP	Ordinal	6	No	20	7	No	Mean
UNESCO	Interval[5]	6	Yes	5	5	No	Mean
UNFPA	Ordinal	4	No	8	8	Yes	Customized
UNICEF	Ordinal	4	Yes	64	6	No	Qualitative
UNIDO	Ordinal	6	Yes	10	1	No	-
UNODC	Ordinal	4	No	58	9	Yes	Not mentioned
WFP	Interval	5	No	9	9	Yes	Quintiles
WIPO	Ordinal	4	No	48	8	Not mentioned	Not mentioned

[4] The author assumed that the scale is of ordinal nature and that numbers 1 to 6 of the QA tool represent a widely-used scale in evaluation such as highly satisfactory to highly unsatisfactory.
[5] UNESCO's scale was considered as interval as its values (0 to 5) were not said to correspond to qualitative answers such as those in the satisfactory scale.

In the remainder of this section, a brief description of QA procedures and methodologies of each entity is mentioned. Noticeable strengths and areas for improvement are highlighted.

FAO

The Office of Evaluation (OED) of the FAO performs biennial ex-post quality assessments (QAs) of project evaluations purposively selected to have a balance regarding geographic and thematic coverage (FAO, 2011a). The FAO assesses the quality of the Terms of Reference (ToR) and the overall evaluation process in addition to that of the draft report (FAO, 2011b). This is a strength of FAO's QA process in relation to others. Whether sections are weighted or not and the method to calculate overall scores of evaluations, ToRs, and draft reports remain unknown.

UN Women

The UN Women's Global Evaluation Report Assessment and Analysis System (GERAAS) has a well-structured methodology. External consultants assess the quality of final evaluation reports on a yearly basis and present it in meta-evaluation reports (UN Women, 2015; UN Women, 2017). Its main strength is that written justifications for scores are provided. This reduces subjectivity in a QA system that uses qualitative judgements to determine overall scores by section and report.

A qualitative aggregation method that is based on, but not determined by, section scores preclude reports that are weak in the major areas (methodology; findings; conclusions and lessons learned; and recommendations) from being highly rated, even if the rest of the sections have high scores. For example, if a report is missing lessons learned, it cannot be rated as satisfactory, even if the rest of the sections score high. The only noticeable downside to this methodology is the higher cost that it might represent.

UNDP

UNDP has made available online a document containing the description of their QA system (UNDP, n.d.) and further guidance for evaluators and evaluation managers to improve quality ex-ante (UNDP, 2009, pp.207-210). Rather than doing a periodical assessment, its system works on a rolling basis, as independent evaluation reports need to be assessed within two weeks of their submission to the Resource Centre. A pre-requisite to a real-time QA process like this is having a well-structured, coordinated evaluation office with a large capacity.

Assessments include a wide-ranging list of evaluation items, divided into six sections or criteria, which are then weighted using a 6-point scale (highly satisfactory to highly unsatisfactory) to calculate the report's overall score. Although the measures of central tendency, with which overall scores by section are calculated, are not described in the guide (UNDP, n.d.), it appears that these are like those at UN Women. Section scores are based on individual ratings at the consultant's discretion. In the calculation of overall scores by report, it looks as if weights are summed by section score, and the rating with the highest sum is given to the whole report.

UNEP

Each one of the evaluation reports at UNEP is assessed using the same 6-point ordinal scale as UNDP (UNEP, 2017; UNEP, n.d.). Only 2 of the 20 items (likelihood of impact and sustainability) are rated on a 6-point scale that goes from highly unlikely to highly likely. The mean of all items is calculated and used as a report's overall score. This could be considered as the main weakness of UNEP's methodology, as means of ordinal data should not be calculated. On the other hand, apart from

assessing 100% of evaluation reports, UNEP's QA system main strength is that the assessment is performed both for the draft report and for the final report. Although this requires a higher investment, it is a practice expected to increase the quality of final reports.

UNESCO

UNESCO's Evaluation Policy states that the Evaluation Office is responsible for conducting meta-evaluations of report quality and synthetic reviews of all completed evaluations (UNESCO, 2014, p. 17). The method used to determine the quality of reports is different from the rest of those analysed here. Five items are assessed using dichotomous variables (yes/no), and the overall score of an evaluation is the sum of the number of present elements in a report (UNESCO, 2016). This can be seen as a limitation, as 'high-quality' reports could instead be 'comprehensive' reports. The presence of lessons learned, for instance, does not mean that these lessons are valuable.

UNFPA

The UNFPA's Evaluation Branch assesses the quality of decentralised country programme evaluation reports using a 4-point ordinal scale (unsatisfactory to very good). Even though each of the scores for the eight sections in the assessment grid (UNFPA, n.d.a) are not the combination of individual item scores, they result from a well-explained set of criteria (UNFPA, n.d.b). After scoring the eight sections, a customised method to calculate the measure of central tendency is used. Simply put, a weight (or multiplying factor) is given to each section, and the score with the highest sum of weights is taken as the overall score.

The 'findings and analysis' section is given a factor of 50 (out of 100), in such a way that most of the time, the score of this section will be the score of the report. The main advantage, which can also be interpreted as a challenge, of UNFPA's methodology is that there are no individual scores. This could make the process of scoring more efficient but also more difficult. Also, even though the measure of central tendency is calculated intuitively, it could be over-relying in the score of one section and, thus, overall scores might be truly representative of the whole report.

UNICEF

The Global Evaluation Reports Oversight System (GEROS) of UNICEF is an organization-wide system that informs managers and stakeholders annually on the quality of evaluation reports, contributing to organisational learning. A 4-point scale of measurement is used to assess a very well-thought set of 64 questions divided into 6 sections (UNICEF, 2013; UNICEF, 2016). Furthermore, some of the 64 questions are considered to be 'key' in such way that they are instrumental in determining the sections' scores.

After this initial assessment, a measure of central tendency is calculated qualitatively by applying a 'test'. Reports are considered good quality if they pass the mentioned test, i.e. if they comply with a set of standards that include credibility, usefulness and evidence-based. This is the main strength of UNICEF's QA system: each score derives from an in-depth analysis of the question, section or report assessed. However, this could also result in a time-consuming and expensive exercise, in comparison with the rest of the QA systems.

UNIDO

The Evaluation Policy of UNIDO states that the Evaluation Office is responsible for maintaining an internal quality assurance system of all its evaluations (UNIDO, 2015a, p.11). Although a document containing the description of its QA methodology could not be found online, an assessment template

was found in one of the evaluation reports (UNIDO, 2015b, p.105). It outlines a set of 10 questions that are scored using the 6-point scale (highly satisfactory to highly unsatisfactory). No evidence was found to say that a measure of central tendency per report is calculated. Overall, the author considers that there is room for improvement in this system, including making assessments independent and the questionnaire more comprehensive.

UNODC

The independent evaluation unit of UNODC contracts external consultants to systematically assess, at the end of each year, the quality of all published evaluation reports and a third of first draft evaluation reports (UNODC, 2016). In this way, the process of revision of draft reports is also evaluated. The 4-point ordinal scale that UN Women uses is also used by UNODC to assess the 58 questions of the nine categories or sections (the 'satisfactory' score is called 'fair' here). The way in which section scores are calculated from individual scores is not mentioned. Neither is mentioned the way of calculating a report's overall score, although the template used by reviewers includes a table of category weightings.

It is possible that these are used in the same way as in UNFPA to determine a measure of central tendency. In this case, the 'findings and analysis' section is also given the largest relevance. In conclusion, this is among the group of 'mature' QA systems, partly because it assesses reports shortly after they are published, it uses a comprehensive checklist, and comprises all published reports. This last good point might not be feasible in agencies with more evaluations (the number of evaluations in UNODC in 2016 was 19).

WFP

In addition to a real-time quality assurance system, WFP has independent quality assessments of all completed evaluations which are reported in the annual evaluation reports (WFP, 2015, p.14). The method used is different from the other UN entities: one score is given to each section, based on several conditions. Each section has a weight from 1 to 10, adding up to 50. As the exact methodology is not described, it is the author's assumption that sections are assessed qualitatively depending on the extent to which they comply with the quality requirements of the section, and that weightings are summed and multiplied by 2 to fit the 0-100 interval scale. Then, a final rating per report is given per quintile (1st quintile = significantly lacking to 5th quintile = excellent). Although sections include a comprehensive set of requirements of good quality, the main challenge of this system is not having individual scores for each of these requirements or items, as section scores might be difficult to determine without individual scores.

WIPO

The Internal Oversight Division of WIPO includes in their evaluation manual a comprehensive and detailed checklist for assessing the quality of evaluation reports that uses a 4-point ordinal scale (WIPO, 2016, pp.39-43). It includes 48 questions divided into eight sections, covering in this way almost every aspect of an evaluation report. However, no information was found about the way in which scores are aggregated, which reports are assessed, the periodicity of assessments, etc.

3. Evaluation quality appraisals in the ILO

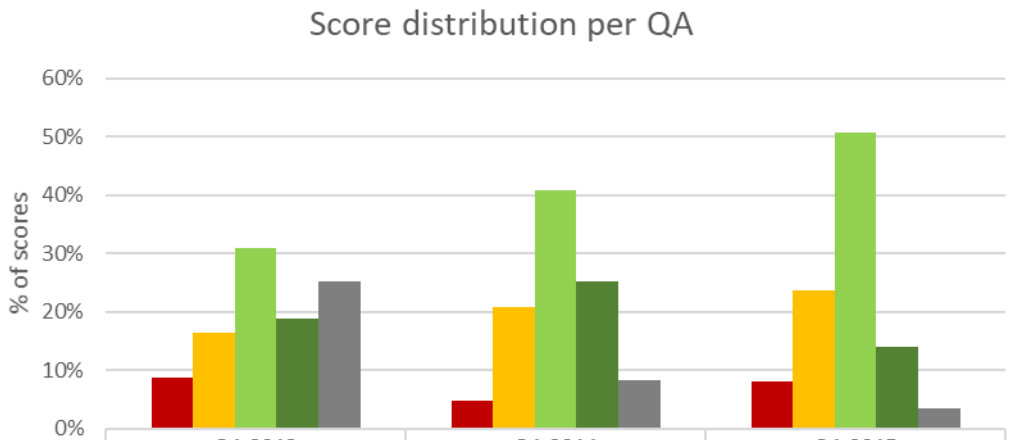
In addition to the comparative analysis described above, to help inform the planned revisions, EVAL conducted an analysis of the last three quality appraisals conducted by the ILO, which covered

evaluation reports between the years 2009 and 2015. The quantitative study used descriptive statistics to identify common characteristics of the lowest and highest scores.
The primary question that EVAL wanted to answer through the analysis was: With all of the initiatives that EVAL has undertaken to improve the quality of evaluation reports (e.g. the Evaluation Manager Certification Programme, updated Evaluation Policy Guidelines, increased budgets to hire better quality consultants, etc.), why did the quality not go up (or down) over time?

The results of the analysis showed that one reason for having persistent levels of quality was the ambiguity in the scoring and analysis of the NA response category. The NA scores were not included in the calculations of central tendency, and so they did not influence the overall score. This means that sections of the report containing only one or two, highly-rated items could score better than those with several, low-rated items. Therefore, the comprehensiveness of the report was not reflected in the overall score.

Analysis showed that the total proportion of NA scores decreased from 25.2% in the 2013 QA to 8.3% and 3.6% in the 2014 and 2015 exercises, respectively (Figure 1). This means that, while the quality of the independent evaluation reports submitted to EVAL over the three appraisal periods may not have gone up, the **comprehensiveness** of the reports certainly did.

Figure 1. Score distribution per QA exercise (2013-2015)



	QA 2013	QA 2014	QA 2015
0	9%	5%	8%
1	16%	21%	24%
2	31%	41%	51%
3	19%	25%	14%
NA	25%	8%	4%

Another reason for the persistent level of quality is that extremely low ratings of certain questions and sections depressed the mean scores. This is because means are sensitive to extremes. The irony of the situation is that extremely low ratings were attenuating the means, however, because the data were on an ordinal scale of measurement, calculating means as a measure of central tendency may not have been an appropriate statistic.

Finally, the analysis identified certain sections, and even specific items, that might have required special attention. In the section on methodology, for example, items dealing with the use of ethical safeguards, evaluation questions, sampling procedures, and the use of Standards and Norms lowered scores. In the recommendations section, items dealing with the time frame, priorities, and resource implications also had a depressive effect. In findings, items dealing with disaggregation of data had lower scores than the rest of the items in the section.

4. Log of changes made to the ILO's QA system

Informed by the review of the quality appraisal systems of similar UN agencies, programmes, and funds and by its analysis of the results of its own past quality appraisals, EVAL made a number of important changes to its quality appraisal system.

First, EVAL revisited the Quality Appraisal tool, itself. A 'demographics' section was added in order to collect data on independent variables that could be used to conduct more advanced statistical analysis. Duplication was eliminated and the tool was simplified. The tool went from having 137 items to only 79. In other words, about 40 per cent of the items were redacted.

How was that possible? For example, the previous tool contained 16 items that collected information on the presence or absence of elements related to the title page. These were consolidated into one item that asked if EVAL's title page template had been used.

The decision to use five response categories (one of which was "NA") was revisited. This was prompted by a concern that the low number of response categories was providing less opportunity for differentiating reports along a continuum of quality.

The number of response categories could have also potentially had statistical implications, as well. With a low number of response categories, the variability in the data was necessarily reduced. EVAL was concerned that this might lower the statistical power of analyses and adversely influence the detection of statistical differences. Therefore, the number of response categories was increased.

The decision to include "NA" among the response categories was also revisited. EVAL carefully examined the items in order to determine in which cases a rating of "NA" would apply. EVAL came to the conclusion that there is nothing in the Quality Appraisal tool that would not be applicable to each and every independent evaluation report. Therefore, "NA" was omitted as a response category.

EVAL also revisited the guidance that was given to consultants regarding the analysis of data. In the past, means (i.e. averages) were calculated as measures of central tendency. However, EVAL came to the determination that the data generated by the QA use an ordinal scale of measurement[6]. This has implications for the statistical techniques that can be used appropriately. Means may not have been an appropriate statistic. Alternative statistical techniques, appropriate to ordinal data, were recommended.

Finally, EVAL recognized that, over the years, different consultant groups have been hired to conduct quality appraisals. Some consultant groups tended to be more rigorous in their ratings of evaluation reports and some were less rigorous. In order to somewhat standardize the rating process, EVAL developed rubrics for assigning scores.

5. Current status

The Quality Appraisal Tool is currently an instrument with a number of annexes and supporting documents. The tool itself contains four sections: 1) a **demographics** section that collects data on variables of each report, such as region, department and year; 2) one that requires reviewers to rate

[6] The data are ordinal (Arora and Trochim, 2013) because the distance between each of the response categories is not equal.

the **quality** of report elements; 3) a **comprehensiveness** check that determines the completeness of reports; and, 4) a section that measures integration of the UN System-wide Action Plan on Gender Equality and the Empowerment of Women (**UN-SWAP**)[7].

The four response categories have been increased to six (highly unsatisfactory to highly satisfactory). The new 6-point scale, which is more widely used across the UN QA systems, allows for a better detection of statistical differences than the 4-point scale because of the possibility of more variance. These response categories are used primarily in the section dealing with quality.

Annex 1 of the Quality Appraisal tool contains recommendations for analysis. The scale of measurement is composed of nominal, ordinal, interval and ratio data. EVAL assumes that the data derived from the comprehensiveness section of the tool is measured on a nominal scale. Therefore, the analyses that would be appropriate would be frequencies and per cents.

The data derived from the quality and the UN-SWAP sections of the tool are assumed to be measured on an ordinal scale. Therefore the analyses that would be appropriate would be frequencies, per cents, mode and median. The use of means is not advised. For the consultants' convenience, a scoring sheet template has also been prepared.

Annex 2 contains a set of scoring rubrics that were developed to somewhat standardize the quality appraisal process. The rubrics present a three-tier structure of components, clustered items (per component) and elements. The rubrics allow an overall rating of items based on the quality of the information provided to the elements or the lack thereof. All components are considered mandatory and thus received an overall score.

The updated QA methodology preserves the strengths of its predecessor: it is comprehensive and clear. The dozens of items ensure that each aspect of the evaluation report is looked at and descriptions of items make the process easy to understand for reviewers.

One thing that could improve further the methodology is weighting sections to calculate overall report scores, as some sections (e.g. recommendations) are more important than others (e.g. project background). This would, of course, require a theoretical rational for determining the relative importance of each section and how much each should be weighted.

6. The possibility of making historical comparisons

The ILO Evaluation Office has expressed a desire to compare the results of future quality appraisals with those conducted in the past. The author is sceptical about the possibility of doing this for reasons of measurement and data equivalence.

[7] It should be noted that the ILO is one of a small number of UN agencies, programmes and funds that relies on independent assessment of the way that gender equality and empowerment of women has been integrated into the ILO's evaluation work.

Measurement Equivalence

Equivalence is said to be achieved when items are measured in such a way that their relationships to underlying concepts are the same across groups[8]. To illustrate this notion, the reader is directed to the table found below that was extracted from the quality appraisal that was conducted by The Evaluation Center on the campus of Western Michigan University (WMU) in 2016.

Table 14. Historical comparison of average scale scores by component area.

Report Section	2014-15	2012-13	2009-11
Overall	1.8	1.9	1.8
Executive Summary	2.0	1.9	1.9
Project Background	2.0	2.1	2.0
Evaluation Background	1.7	1.9	1.9
Evaluation Methodology	1.5	1.6	1.7
Findings	1.8	1.9	1.7
Conclusions	1.8	2.1	2.1
Recommendations	1.7	1.8	1.3
Lessons Learned	1.8	1.8	1.9
Good Practices	1.9	1.5	1.9
Formal Elements	1.9	n.a	1.9

To construct this table, WMU created indices by taking an average of the averages of all the items in each report section for each year that the quality appraisal was conducted. Because the Quality Appraisal Tool remained substantially unchanged and the procedures for conducting the appraisal were the same, this comparison seems methodologically defensible and one could presume measurement equivalence.

However, given (1) the radical changes to the Quality Appraisal Tool (e.g. reduction of 40 per cent of the items) and (2) the new procedures for conducting an appraisal (e.g. new scoring rubrics) that were described above, the author believes that measurement equivalence cannot be presumed.

Data Equivalence

Not only would measurement not be equivalent, the data would also not be equivalent. The definition above can be usefully modified as follows: data equivalence is said to be achieved when the relationship among responses to items and underlying concepts are the same across groups.

Because (1) the composition of items that were used to create the indices for each report section will be different, (2) the data will be assumed to be measured on a different scale of measurement, (3) the number of response categories will increase to six, and (4) different statistics will be calculated (i.e. non-parametric rather than parametric statistics), the author believes that data equivalence also cannot be presumed.

Historical comparisons, in the absence of statistical analysis to demonstrate measurement and data equivalence, run the risk of diminishing the usefulness of findings and, in the worst case scenario, rendering the results totally meaningless.

[8] Pendergast, L., von der Embse, N., Kilgus, S. & Eklund, K. (2017). Measurement equivalence: A non-technical primer on categorical multi-group confirmatory factor analysis in school psychology. Journal of School Psychology. (60) 65-82.

7. Lessons learned

It could reasonably be asked; what was the ILO's takeaway from this study? Three lessons learned and five good practices emerged from the analysis.

- **Scales of measurement should be harmonised across the UN**
  Although 10 of the 12 agencies are using 4- or 6-point ordinal scales, these do not have the same scores. Some agencies have opted for commonly used scales: ILO, UNDP, UNEP, and UNIDO use the 6-point satisfactory scale, and three other (UN Women, UNFPA, UNODC) use a 4-point scale that ranges from 'unsatisfactory' to 'very good'. The scales used by UNICEF and WIPO were not found in other agencies, and FAO's scores' meanings were not found. The author considers that evaluation in the UN system would benefit from harmonisation of QA scales. This would both help external reviewers and enable comparisons across entities.

- **The calculation of measures of central tendency (overall scores) needs to be tailored to the needs of the organisation**
  One of the main features observed was the existence of large differences among UN entities in the calculation of overall scores by report and section. These vary from qualitative methods that rely on the professional judgement of the evaluator to systematised quantitative methods (means, medians, modes and quintiles). Some methods assign weights to each section, some incorporate 'intrinsic weights', and some assign the same value to each section. All these decisions should be made considering the priorities and needs of each entity.

- **The presence of an element does not mean that an evaluation is of good quality**
  As experienced by the ILO, there needs to be a clear distinction between a report's comprehensiveness and its quality. The presence of an element in a report is a necessary, but not sufficient, condition for being of high quality.

8. Good practices

- **Assessing the quality of ToR and draft reports**
  In addition to performing quality assessments on final evaluation reports, some entities assess the quality of terms of reference and draft reports. This practice is likely to promote real-time organisational learning and to improve the quality of final evaluation reports.

- **Using comprehensive questionnaires for assessments**
  The use of questionnaires or checklists that include all the relevant items leads to a better QA exercise. Not only does it ensure that every aspect of the report is reviewed, but it promotes the improvement of overall quality through the correction of specific issues within report sections. Having more (but not too many) questions enables analyses that detect the exact aspect on which quality could be higher.

- **Leveraging quantitative methods for the analysis of scores**
  Large efforts are invested to put quality scores in 4 or 6-point scales. A quantitative method of analysis should be used to analyse this valuable data. Trends, areas for improvement, and good practices can be identified using distributions, medians, modes, and other calculations. Today, at least seven entities (ILO, UN Women, UNDP, UNESCO, UNFPA,

UNICEF, and UNODC) use distributions to analyse results. Others should follow their lead and leverage the use of quantitative methods for comparison.

- **Balancing the selection of evaluation reports assessed**
  Certain agencies have succeeded at assessing all their evaluations. However, it is recognized that some entities do not have the capacity to assess 100% of the reports on a given period. Therefore, the selection of evaluation reports needs to take into account a geographical and thematic balance. This could be achieved, for instance, by doing simple or stratified randomization.

- **Contracting external consultants to perform QAs**
  The analysis showed that quality assessments are not always performed by external consultants. This practice is important to ensure the independence of the exercise.

9. Conclusions

The first objective of this THINK Piece was to review the ILO's methodology in order to assess the quality of project-level evaluations and to provide an explained snapshot of its recent results. More specifically, the latter tried to explain why the same level of 'average quality' persisted, rather than increased, over the last three QA exercises (covering evaluations from 2009 to 2015).

Two explanations were found. On the one hand, specific questions and sections scored consistently low, and this kept report average scores from rising. More focus and guidance for evaluators is suggested in these areas. On the other hand, a fact that was not being looked at was that report comprehensiveness was increasing. So, even if the quality of existing elements did not increase, reports became more complete likely because of more complete guidance, training of evaluation managers and stricter real-time quality control. This can also be considered an improvement of overall report quality.

An additional point needs to be made. The term 'average quality' was intentionally used to point to the fact that what was being looked at by previous reviewers was a mean/average of ordinal data. Since means should not be calculated for such type of data, EVAL changed its QA methodology.

EVAL's 2017 'update' of its QA methodology amends the two main issues found in the quantitative study. First, its inclusion of a 'comprehensiveness' dimension ensures that both aspects of quality are looked at. Second, medians and modes substitute means as measures of central tendency. Moreover, the 4-point scale is now a 6-point scale that is both recommended and commonly used in the UN system (UNDP, UNEP, UNIDO).

The new scale is more sensitive to changes in overall scores. In conclusion, EVAL made relevant changes to its QA methodology, which can be used by UN entities that wish to improve their QA system.

References

Arora, K. and Trochim, W. (2013) Rating Systems in International Evaluation. ILO-EVAL THINK Piece No. 3. Available at: http://www.ilo.org/wcmsp5/groups/public/---ed_mas/---eval/documents/publication/wcms_202430.pdf

FAO (2011a) Quality Assurance Framework for evaluation in FAO. Office of Evaluation. Available at: http://www.fao.org/fileadmin/user_upload/oed/docs/Evaluation_Docs/Guidlines/Framework_for_OED_Quality_Assurance_Framework.pdf

FAO (2011b) Tools for Quality Assurance of OED evaluation terms of reference and reports. Office of Evaluation. Available at: http://www.fao.org/fileadmin/user_upload/oed/docs/Evaluation_Docs/Guidlines/OED_QA_tools_final_version.xlsx

Friedman, J. and Blight, N. (2014) External Quality Appraisal: Implications for evaluation quality and utilization. i-eval THINK Piece No. 8. Available at: http://www.ilo.org/eval/newsletter-and-think-pieces/WCMS_329163/lang--en/index.htm

ILO (2005) Evaluation. A new policy and strategic framework for evaluation at the ILO. GB.294/PFA/8/4. Available at: http://www.ilo.org/public/english/standards/relm/gb/docs/gb294/pdf/pfa-8-4.pdf

ILO (2011) Results-based strategies 2011–15: Evaluation strategy – Strengthening the use of evaluations. GB.310/PFA/4/1(Rev.). Available at: http://www.ilo.org/wcmsp5/groups/public/---ed_norm/---relconf/documents/meetingdocument/wcms_152025.pdf

ILO (2017) ILO Evaluation Policy. GB.331/PFA/8. Available at: http://www.ilo.org/wcmsp5/groups/public/---ed_mas/---eval/documents/policy/wcms_603265.pdf

ILO-EVAL (2014) Checklist 6: Rating the Quality of Evaluation Reports. i-eval Resource Kit. Available at: http://www.ilo.org/wcmsp5/groups/public/---ed_mas/---eval/documents/publication/wcms_165968.pdf

ILO-EVAL (2017a) Revised Appraisal Tool and Recommendations for Analysis. Not publicly available.

ILO-EVAL (2017b) Independent evaluation of the ILO's Evaluation Function, Main Report. ImpactReady and Lattanzio. Available at: http://www.ilo.org/wcmsp5/groups/public/---ed_mas/---eval/documents/publication/wcms_545949.pdf

Prom-Jackson, S. and Bartsiotas, G. (2014) Analysis of the Evaluation Function in the United Nations System. JIU. Geneva. Available at: https://www.unjiu.org/en/reports-notes/JIU%20Products/JIU_REP_2014_6_English.pdf

Robertson, K. and Schroter, D. (2014) Leveraging appraisal findings to improve evaluation quality. Ieval THINK Piece No. 4. Available at: http://www.ilo.org/eval/newsletter-and-think-pieces/WCMS_237914/lang--en/index.htm

UN Women (2015) Global Evaluation Report Assessment and Analysis System (GERAAS). Available at: http://www.unwomen.org/-

/media/headquarters/attachments/sections/about%20us/evaluation/evaluation-geraasmethodology-en.pdf

UN Women (2017) Global Evaluation Report Assessment and Analysis System. Meta Evaluation Report 2016. Available at: https://gate.unwomen.org/EvaluationDocument/Download?evaluationDocumentID=9057

UNDP (2009) HANDBOOK ON PLANNING, MONITORING AND EVALUATING FOR DEVELOPMENT RESULTS. Available at: http://web.undp.org/evaluation/handbook/documents/english/pme-handbook.pdf

UNDP (n.d.) Quality Assessment System for Decentralized Evaluation Reports. Evaluation Office. Available at: http://web.undp.org/evaluation/documents/guidance/UNDP-Quality-Assessment-System-for-Decentralized-Evaluation.pdf

UNEG (2016) Norms and Standards for Evaluation Available at: http://www.unevaluation.org/document/detail/1914

UNEP (2017) Quality Assessment of the Evaluation Report. Evaluation Office. Available at: http://wedocs.unep.org/bitstream/handle/20.500.11822/7108/18.%20Template%20for%20the%20Assessment%20of%20the%20Quality%20of%20the%20Evaluation%20Report.pdf?sequence=1&isAllowed=y

UNEP (n.d.) Quality Assurance of Evaluation Office Reports. Available at: http://www.unep.org/evaluation/policy-standards/quality-assurance
UNESCO (2014) Evaluation Policy 2014-2021. 196 EX/24.INF. Available at: http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/IOS/images/UNESCO_Evaluation_Policy_EN.pdf

UNESCO (2016) Synthetic Review of Evaluations in the UNESCO System. Evaluation Office.

IOS/EVS/PI/148 REV.2. Available at: http://unesdoc.unesco.org/images/0024/002443/244354e.pdf

UNFPA (n.d.a) Evaluation Quality Assessment Grid Template. Available at: http://web2.unfpa.org/public/about/oversight/evaluations/templates.unfpa

UNFPA (n.d.b) Explanatory notes for UNFPA Evaluation Quality Assessment Grid. Available at: https://www.unfpa.org/sites/default/files/admin-resource/EQA%20Explanatory%20Note%20for%20Country%20Offices.pdf

UNICEF (2013) Global Evaluation Reports Oversight System (GEROS). Evaluation Office. Available at: https://www.unicef.org/evaluation/files/GEROS_Methodology_v7.pdf

UNICEF (2016) UNICEF GEROS Meta-Analysis 2015, An independent review of UNICEF evaluation report quality and trends, 2009-2015. Available at: https://www.unicef.org/evaldatabase/files/UNICEF_GEROS_Meta_Analysis_v2_1(print).pdf

UNIDO (2015a) DIRECTOR GENERAL'S BULLETIN: EVALUATION POLICY. UNIDO/DGB/(M).98/Rev.1. Available at: http://www.unido.org/fileadmin/user_media_upgrade/Resources/Evaluation/UNIDO_Evaluation_Policy_UNIDO-DGB-M-98-Rev-1_150319.pdf

UNIDO (2015b) Independent final evaluation of the National Cleaner Production Programme – Republic of Moldova. UE/MOL/11/002, SAP ID 104143. Available at: http://www.unido.org/fileadmin/user_media_upgrade/Resources/Evaluation/UEMOL11002-104143_NCPP_EvalRep-F_151020_01.pdf

UNODC (2016) Independent Quality Assessment of UNODC Evaluation Reports 2016. Synthesis Report. Available at: https://www.unodc.org/documents/evaluation/EvaluationQualityAssessments/UNODC_Independent_Evaluation_Quality_Assessment_2016.pdf

UN-OIOS/IED (2014) Inspection and Evaluation Manual. Available at: https://oios.un.org/resources/2015/01/OIOS-IED_Manual.pdf

Watts, B. (2016) Quality assessments of ILO project evaluations: What are the next steps to better evaluations? i-eval THINK Piece No. 10. Available at: http://www.ilo.org/eval/newsletter-and-think-pieces/WCMS_329163/lang--en/index.htm

Watts, B., Coryn, C., Feikowsky, E., Wilson, L. and Mateu, P. (2015) Independent Quality Appraisal of ILO Project Evaluations. The Evaluation Center, Western Michigan University.

WFP (2015) EVALUATION POLICY (2016–2021). WFP/EB.2/2015/4-A/Rev.1. Available at: http://documents.wfp.org/stellent/groups/public/documents/eb/wfp277482.pdf?_ga=2.204858198.790843628.1498584765-352740954.1494428130

WIPO (2016) Evaluation Manual. Available at: http://www.wipo.int/export/sites/www/about-wipo/en/oversight/iaod/evaluation/pdf/evaluation_manual.pdf