GEROS – Global Meta-Evaluation Report 2014

Final Report

September 2015

Universalia
245 Victoria Avenue, Suite 200
Westmount, Quebec
Canada H3Z 2M6
www.universalia.com

Executive Summary

The purpose of UNICEF's Global Evaluation Quality Assurance System (GEROS) is to ensure that UNICEF's evaluations uphold the high quality standards set for them. The GEROS review process and the information provided in annual meta-evaluation reports help UNICEF to monitor its progress, identify its strengths and become aware of its areas for improvement with regard to evaluation.

The four main objectives of GEROS are to: (i) provide senior managers with a clear and independent assessment of the quality and usefulness of evaluation reports; (ii) strengthen internal evaluation capacity, through practical feedback on how to improve future evaluations; (iii) contribute to corporate knowledge management and organizational learning, making available good quality evaluations; and (iv) report to the Executive Board on the quality of evaluation reports.

Each of UNICEF's regional offices is responsible for submitting completed evaluations to the GEROS process. These include a diverse range of evaluations which differ in terms of thematic area, geographic coverage, management approach, and other elements.

Methodology

Reports completed during the 2014 GEROS review cycle were reviewed over a five-month period (mid-January to early June 2015). In total, 69 reports were analysed using the GEROS assessment tool, which contains 58 standards for evaluation reports established by UNICEF. In the GEROS review process, each question and section is given a rating based on a colour-coded four-point performance scorecard that goes from Outstanding/Best Practice to Unsatisfactory. For each of the section and sub-section ratings, a narrative justification is also provided and supported by evidence from the report (e.g. examples and page or section numbers).

Profile of reports reviewed in 2014

Most reports were submitted by the Central and Eastern European/Commonwealth of Independent States Regional Office (CEE/CIS), the South Asia Regional Office (ROSA), and the East and Southern Africa Regional Office (ESARO). The East Asia and Pacific Regional Office (EAPRO), the Latin America and Caribbean Regional Office (LACRO), and the Middle East and North Africa Regional Office (MENARO) submitted far fewer reports than in 2013, with five or less reports per regional office in 2014.

Most reports presented the results of either programme or project evaluations and all evaluation reports reviewed this year were carried out by independent external evaluators.

This meta-analysis report presents the analysis of aggregate ratings from the 2014 review cycle. Reports with a rating of Highly Satisfactory or Outstanding/Best Practice, are considered to be "quality" reports. Data from recent GEROS cycles (i.e. from 2010 to 2013) has also been included for comparison purposes.

Overarching conclusions, recommendations and lessons learned are presented below.

Conclusions

The quality of UNICEF's evaluation reports continued to increase through 2014, but only moderately. The upward trend in report quality maintained itself for another year, as good quality ratings reached 74% (from 69% in 2013 and 62% in 2012). This indicates that most reports do an adequate job of meeting GEROS standards.

National level evaluation reports increased in quality compared to previous years (from 50% in 2013 to 63% in 2014). Project-level evaluations increased significantly in quality when compared to 2013 (from 58% in 2013 to 84% in 2014). In terms of quality by areas, WASH evaluations had the lowest quality rating, at only 29%. Education, on the other hand, had the highest quality rating at 82%. In terms of evaluation management, i.e. control and oversight of evaluation decisions, the quality of UNICEF-managed evaluations continued to improve (from 74% in 2013 to 80% in 2014). In 2014, the highest rated sections of evaluation reports were "Methodology" with 71% of quality ratings and "Findings and Conclusions" with 74% of quality ratings. The lowest rated section of evaluation reports was the one on "Recommendations and lessons learned" with only 52% of quality ratings.

The contribution of reports to the GEROS process by some of UNICEF's regional and country offices appears to be dropping. In 2014, the number of reports reviewed dropped considerably, decreasing from 96 to 69. This drop reflects fewer submissions from WCARO, LACRO, ESARO, and MENARO. There are several factors that may affect the number of evaluation reports submitted in a given GEROS cycle, and no firm conclusions can be drawn to explain this decrease. In some cases, fewer evaluations may be conducted in one year compared to another. This is sometimes related to low implementation rates, in which projects or programmes have not progressed enough to justify an evaluation. Evaluations may be planned but not all executed due to issues related to budget, lack of priority, or staff turnover. Finally, in some cases low quality evaluations may be reclassified as a review and subsequently not submitted to GEROS. It is unclear which of these factors had the greatest effect on the lower number of evaluation reports in 2014.

Reports continue to demonstrate similar shortcomings found in 2013. While good quality ratings for most of the GEROS standards improved, some standards remained particularly weak. The description of the evaluated object's theory of change, key stakeholder contributions, and ethical considerations and safeguards was often missing or incomplete. These can be easily addressed by systematically requiring evaluators to include these elements. In other cases, a significant improvement will require different approaches to an evaluation, as in conducting cost analysis or better incorporating gender, human rights from start to finish of an evaluation assignment, including in the design of the Terms of Reference. Reports were also frequently missing a clear rationale for the use of identified OECD/DAC evaluation criteria and recommendations that are targeted to specific stakeholders, and clear lessons learned. It would also appear that many evaluators and/or evaluation managers do not fully understand UNICEF's definition or expectations around lessons learned, such as the need to have lessons that can be generalized to indicate wider relevance to other contexts and situations.

Recommendations

UNICEF should examine whether the increase in quality of evaluation reports has resulted in senior managers having greater confidence in evaluation reports.

Given that evaluation quality appears to be growing, efforts should be made to analyse how the GEROS system contributes to the increased use and value of UNICEF's evaluation function, as a whole. More specifically, UNICEF should see if, as a result of the increased reporting quality noted, senior managers now have greater confidence in the reports produced and whether the information obtained is useable.

Within its decentralised evaluation strategy, UNICEF should continue to build its own regional/country office evaluation capacities and national capacities to conduct relevant types of evaluations.

While good foundations have been established, there is still a need to ensure that targeted, focused support is provided to build UNICEF regional and country office capacity and national evaluation capacities. This capacity support should take into account types of evaluations that are currently not frequently conducted, including: evaluation of strategies, country-level evaluations, and country-led evaluations.

Special efforts should be made to strengthen certain aspects of evaluation reports that have been consistently weak in the past few years.

Over the past few GEROS cycles, performance on some evaluation components has been consistently weak (e.g. theory of change, cost analysis, identification of lessons learned). In order to address these weaknesses, UNICEF should focus further efforts and attention on improved training and guidance around these standards in particular.

UNICEF should continue to update and systematically communicate its requirements for evaluation reports across its entire evaluation oversight/management system. These updates should take into account evolving standards for evaluation in the UN System.

UNICEF should ensure that its standards and criteria are clear and up-to-date across all evaluation management levels. This includes reviewing and clarifying the GEROS framework and standards (which UNICEF does periodically) as well as the systematic integration of evaluation priorities and standards in all TORs. This should also apply to the UN-SWAP requirements for gender-responsive evaluation, which have not been disseminated widely.

As part of the periodic review of GEROS , UNICEF should consider revising the rating scale and several elements of the GEROS template in order to ensure greater precision in the messages that are provided about evaluation quality and the characteristics of evaluation reports, and to create more efficiency in applying the template.

For the 2012 cycle, after three years of implementing GEROS, UNICEF changed the rating scale. The rating of "Almost Satisfactory" was changed to "Mostly Satisfactory", while the traffic light colour (yellow) associated with the rating remained the same. The change in the rating created a larger gap between "Unsatisfactory" and "Mostly Satisfactory" ratings, which creates challenges in making the judgment about how to rate particular questions or standards in the template. Appendix III provides further comments on the scale and identifies other areas where greater clarity or efficiency can be introduced in applying the evaluation report classification and quality standards, as described in the GEROS template.

Lessons Learned

Clear and systematic communication of evaluation standards and priorities favours the effective alignment of evaluations with UNICEF standards, from the outset (i.e. TORs stage).

In order to ensure that evaluations respond to UNICEF's expectations and priorities, evaluation TORs should clearly identify the standards by which reports will be judged in the GEROS process.

While common standards help improve evaluation quality, quality assurance systems such as GEROS should provide sufficient flexibility to account for different types of evaluations.

A balance should be found between common standards and flexibility, to take into account the different types of evaluations produced each year within UNICEF. Currently, the GEROS template used does not always easily adapt itself to specific types of evaluation reports (e.g. impact, separately published case study evaluations which have been completed as part of a broader evaluation).

In a decentralized system, compliance with quality assurance systems such as GEROS is affected by incentives, available resources, and the perception of relevance.

For full participation in a process such as GEROS, the relevance of the process must be clear to the UNICEF regional stakeholders. Part of maintaining relevance is ensuring that the system is continuously updated to reflect changing corporate and UN-system expectations and priorities. In addition, UNICEF regions should have equitable access to the resources (human and other) to support quality enhancement to evaluations in their regions.

Quality assurance systems such as GEROS need to strike a balance between consistent application over a period of time (which allows for comparison) and making major adjustments in order to improve utility and reflect changes in the environment.

Updating and changing the GEROS template every year is not necessary and would be time consuming. It is however necessary to have a formal and consultative review every two to three years in order to keep the process relevant and to update the GEROS template as required in light of new developments or requirements.

Acronyms

CEE/CIS	Central and Eastern Europe/Commonwealth of Independent States (Regional Office)
COs	Country Offices
EAPRO	East Asia and Pacific Regional Office
EO	Evaluation Office
ESARO	East and Southern Africa Regional Office
GEROS	Global Evaluation Report Oversight System
HQ	Headquarters
HRBAP	Human Rights Based Approach to Programming
LACRO	Latin America and Caribbean Regional Office
M&E	Monitoring and Evaluation
MENARO	Middle East and North Africa Regional Office
MTSP	Medium Term Strategic Plan
N/A	Not Applicable
OECD/DAC	Organisation for Economic Co-operation and Development/Development Assistance Committee
RBM	Results-based Management
ROs	Regional Offices
ROSA	Regional Office of South Asia
RTE	Real-time evaluation
SPOA	Strategic Plan Objective Area
SWAP	System-wide Action Plan
ToC	Theory of Change
TORs	Terms of Reference
UN	United Nations
UNDAF	United Nations Development Assistance Framework
UNEG	United Nations Evaluation Group
UNICEF	United Nations Children's Fund
WASH	Water, Sanitation and Hygiene
WCARO	West and Central Africa Regional Office

Contents

1 Introduction 1

2 GEROS Background 2

3 Purpose, Objectives and Scope of GEROS 3

4 Methodology 4

4.1 Review of Evaluation Reports 4
4.2 Meta-Evaluation 6
4.3 Changes Made From Previous Years 7
4.4 Limitations 7

5 Findings 9

5.1 Overall Ratings and Feedback 9
5.2 Overall Regional Trends 11
5.3 Trends by Type and Scope of Evaluation 12
5.4 Trends by Quality Assessment Category 17
5.4.1 Overall Trends All Sections 17
5.4.2 Section A: Object of the Evaluation 18
5.4.3 Section B: Evaluation Purpose, Objectives and Scope 20
5.4.4 Section C: Evaluation Methodology, Gender, Human Rights and Equity 22
5.4.5 Section D: Findings and Conclusions 26
5.4.6 Section E: Recommendations and Lessons Learned 28
5.4.7 Section F: Report Structure, Logic and Clarity 30
5.4.8 Coherence 32
5.4.9 Addressing the gaps in Evaluation Reports 32
5.5 Examples of Good Practices 33

6 Conclusions 35

7 Recommendations 37

8 Lessons Learned 40

Exhibits

Exhibit 4.1 Performance Scorecard 5
Exhibit 5.1 Reports Reviewed per Region per Year 9
Exhibit 5.2 Good Quality Evaluation Reports over Time (2009-2014) 10
Exhibit 5.3 Overall Ratings for 2014 10
Exhibit 5.4 Proportion of Good Quality Reports per Region 12
Exhibit 5.5 Proportion of Reports Reviewed and Good Quality Reports by Purpose (2014) 14
Exhibit 5.6 Changes in MTSP and SPOA Correspondence categories 14
Exhibit 5.7 Proportion of Reports Reviewed and Good Quality Reports by SPOA Correspondence 15
Exhibit 5.8 Proportion of Reports by Level of Independence (2009-2014) 16
Exhibit 5.9 Good Quality Ratings per Section – Year by Year Progression (2011-2014) 18
Exhibit 5.10 Section A Ratings and Comparison (2011-2014) 18
Exhibit 5.11 Sub-Section Ratings: Object of the Evaluation 20
Exhibit 5.12 Section B Ratings and Comparison (2011-2014) 21
Exhibit 5.13 Sub-Section Ratings: Purpose, Objectives and Scope of the Evaluation 22
Exhibit 5.14 Section C Ratings and Comparison (2011-2014) 22
Exhibit 5.15 Sub-Section Ratings: Methodology, Gender, Human Rights, and Equity 24
Exhibit 5.16 Inclusion of Human Rights, Gender, and Equity: Good Quality Ratings Year by Year 25
Exhibit 5.17 Section D Ratings and Comparison (2011-2014) 26
Exhibit 5.18 Sub-Section Ratings: Findings and Conclusions 27
Exhibit 5.19 Section E Ratings and Comparison with (2011-2014) 28
Exhibit 5.20 Sub-Section Ratings: Recommendations and Lessons Learned 30
Exhibit 5.21 Section F Ratings and Comparison (2011-2014) 30
Exhibit 5.22 Sub-Section Ratings: Report Structure, Logic and Clarity 31

Appendices

Appendix I List of Findings 41
Appendix II List of Recommendations 43
Appendix III Detailed Comments on GEROS Framework 44
Appendix IV Terms of Reference 50
Appendix V List of Reports Assessed 53
Appendix VI GEROS Assessment Tool 60
Appendix VII Regional Ratings Graphs 76
Appendix VIII UNICEF, Country-led and Jointly Managed Evaluations by Region 80
Appendix IX Overall Ratings Graphs by Report Typology 85
Appendix X Sub-Section Ratings 89
Appendix XI Analysis Table for 2014 Reports 93
Appendix XII Clarification of Criteria for Completing GEROS Reviews 95
Appendix XIII Special Cases in the 2014 GEROS Exercise 100

1 Introduction

The Universalia Management Group Limited (Universalia) is pleased to submit to UNICEF the final Global Meta-Evaluation Report on 2014 evaluation reports. As part of the Global Evaluation Reports Oversight System (GEROS), the report aims to highlight important trends, weaknesses, strengths, lessons learned and good practices drawn from our review of the 69 final evaluation reports submitted in the 2014 cycle. Additionally, the GEROS process and template is analysed, and possible improvements are suggested in order to foster continuous learning within UNICEF.

In this report, quantitative analyses of the documents reviewed and average ratings are complemented by qualitative assessments of strengths and areas for improvement. Good practices are identified and suggestions for improvements are provided, where relevant. The main conclusions, recommendations and lessons learned derived from this analysis are presented toward the end of the report.

The report is structured as follows:

- Section 2 – GEROS Background
- Section 3 – Purpose, Objectives and Scope of GEROS
- Section 4 – Methodology
- Section 5 – Findings
- Section 6 – Conclusions
- Section 7 – Recommendations
- Section 8 – Lessons Learned

The Appendices are composed of important supporting documentation, notably the Terms of Reference (Appendix IV), GEROS assessment tool (Appendix VI) and additional graphs showing key trends (Appendices VII to XI).

2 GEROS Background

The 2013 Evaluation Policy defines evaluation as a shared function within UNICEF, with key roles distributed across senior leaders and oversight bodies, heads of offices, technical evaluation staff and sectoral programme staff. UNICEF's Evaluation Office (EO) and Regional Offices (ROs) collaborate in order to strengthen the organisation's evaluation function. Both levels play quality assurance roles, aiming to ensure that the evaluations managed or commissioned by UNICEF uphold the high quality standards set for them.

Because UNICEF is decentralised in nature, its evaluations are generally commissioned and managed at the Country Office level. On one hand, such an arrangement helps ensure that report analyses remain highly focused on the national context, but on the other, this decentralised system makes it difficult to maintain uniform quality, high credibility and utility of the evaluations produced organisation-wide. Recognising that M&E functions and practices tend to vary across regions and Country Offices, UNICEF created GEROS as a tool to assess quality of evaluations and inform further development of the organisation's evaluation function. After six years of implementation, some small adjustments were made (see Section 4.3), but the template used to assess 2014 reports is virtually identical compared to the previous year. The template has been used in a more flexible way when applied to impact evaluations.

The GEROS process and annual meta-evaluations provides an important source of information for UNICEF to monitor its progress, identify its strengths and become aware of its areas for improvement with regards to evaluations conducted around the world on its behalf.

3 Purpose, Objectives and Scope of GEROS

The four main objectives of GEROS are as follows:

1) To provide senior managers in the UNICEF HQ, Regional Offices and Country Offices with a clear and succinct independent assessment of individual evaluation reports;
2) To strengthen internal evaluation capacity by providing feedback to commissioning offices with practical recommendations to improve future evaluations and to inform their own assessment of the performance of external consultants who might be hired for future evaluations;
3) To report on the quality of evaluation reports through a review and assessment process whose results are communicated to senior management and inclusion of this information in the Global Evaluation Database; and
4) To contribute to the EO's corporate knowledge management and organisational learning, by providing the evidence base for a meta-analysis of good quality reports.

These objectives constitute the foundation of the GEROS process, an organisation-wide system that is strongly anchored in the United Nations Evaluation Group (UNEG) standards and other UNICEF-specific standards, notably the importance of considering gender-responsiveness, equity and human rights-based approaches in the reporting process. Country offices (COs) and ROs can also take part in additional quality-assurance mechanisms that are designed to assess and improve the quality of Terms of Reference (TORs), inception reports and draft evaluation reports, before they are submitted to the final evaluation review process.

In order to ensure the integrity and reliability of the process, an independent consulting firm conducts the GEROS reviews and produces the meta-analysis, each year. The criteria of assessment, laid out in a detailed assessment tool (Appendix VI), are the basis for ensuring in-depth analysis of the content and quality of the reports submitted through COs, ROs and HQ.

4 Methodology

For 2014,[1] as in previous cycles, the GEROS review process was conducted in three main phases. The process was initiated with a review-team briefing in which the GEROS background and objectives were explained, the assessment tool was presented, key considerations and clarifications were provided, and lessons learned from the 2013 cycle were shared. The briefing was followed by the methodical assessment of each UNICEF evaluation submitted in 2014, and the data obtained was analysed in order to produce the meta-evaluation. The methodology for the latter two phases is described in the following sections.

4.1 Review of Evaluation Reports

To ensure the validity, credibility and uniformity of the assessment, the reviews were themselves conducted in two phases: first, each report was systematically evaluated according to the GEROS assessment tool, and reviews underwent a quality assurance process, both internally and with UNICEF.

Review Process and Evaluation Tool

The Universalia review team was granted a five-month period to review the evaluation reports uploaded to the UNICEF Global Evaluation Database for 2014. Though UNICEF's EO first screened the reports to ensure that submissions were in fact evaluations,[2] an additional screening took place internally upon beginning each review: the top section of the assessment tool serves to categorise each report according to its geographic coverage, managerial oversight, purpose, level of change sought, Strategic Plan Objective Area (SPOA) correspondence, level of independence, and approach (i.e. summative, formative or both). In the end, all evaluation reports uploaded to the database in 2014, for a total of 69 reports,[3] were reviewed from mid-January to early June 2015 (for a full list, see Appendix V).

[1] The review process begins immediately after all reports are submitted for a given year. Thus, reports dated 2014 are reviewed in the 2015 calendar year, but are considered part of the 2014 review cycle. As noted in Section 1, some 2015 evaluations were included in the 2014 report population.

[2] Because the Global Evaluation Database contains surveys, research and baseline studies of different types, in addition to evaluations, this screening process helped ensure that only the appropriate reports were included in the GEROS process. UNICEF defines an evaluation as a "judgment [on] the relevance, appropriateness, effectiveness, efficiency, impact and sustainability of development efforts, based on agreed criteria and benchmarks among key partners and stakeholders. It involves a rigorous, systematic and objective process in the design, analysis and interpretation of information to answer specific questions. It provides assessments of what works and why, highlights intended and unintended results, and provides strategic lessons to guide decision-makers and inform stakeholders" (UNICEF, 2011).

[3] It is important to note that any evaluation conducted by the service provider for GEROS (Universalia Management Group) was rated by another external company; in 2014 there was only one of these cases.

The reports received were assigned to reviewers, where possible, according to thematic expertise (e.g. gender) and linguistic abilities (i.e. each review was completed in the same language as the report itself). The GEROS assessment tool used in the review process contains 58 questions derived from the UNICEF-adapted UNEG evaluation reports standards.[4] These questions are distributed across six main sections, as follows:

- Object of the evaluation;
- Evaluation purpose, objectives and scope;
- Evaluation methodology, gender, and human rights;
- Findings and conclusions;
- Recommendations and lessons learned; and
- Structure, logic, and clarity of the report.

These overarching sections contain 22 sub-sections structured around fundamental areas of focus (e.g. methodological robustness, soundness of findings, depth and value-added of the conclusions). In the GEROS review process, each question and section were given a rating based on a colour-coded four-point performance scorecard (as demonstrated in Exhibit 4.1 below). For each rating and section, a narrative justification was provided and supported by evidence, examples and page/section numbers.

Exhibit 4.1 Performance Scorecard

Colour Coding	CC	Dark Green	Green	Amber	Red	White
	Questions	Outstanding Best Practice	Highly satisfactory	Mostly Satisfactory	Unsatisfactory	Not Applicable

A qualitative approach was taken to each report assessment and served to generate targeted, relevant analysis that took into account each report's specificities and context, in addition to providing constructive feedback for future reports. As shown above, the sections and overall report received grades of: "Outstanding, Best Practice," "Highly Satisfactory," "Mostly Satisfactory," or "Unsatisfactory." Though it was used sparingly, an "N/A" option was also available to account for special cases where certain questions were not relevant. In such instances, the reviewers were expected to explain why the nature of the report made integrating a given component impossible.

Next to the ratings, a narrative was provided to justify the grade attributed to the sub-sections and overarching sections. For its part, the "Constructive Feedback" column in each main section was used to highlight either areas for improvement (including examples and justification) or areas of great strength/best practice that should be maintained in future reports. After each main section, a few sentences summarising key ratings and comments were included under "Executive Feedback." (See Appendix VI for an overview of the scorecard.)

[4] The components of the GEROS assessment tool are anchored in the eight UNICEF-adapted UNEG evaluation reports standards, which consist of: (i) the report structure; (ii) the object of evaluation; (iii) the evaluation purpose, objective(s) and scope; (iv) the evaluation methodology; (v) findings; (vi) conclusions and lessons learned; (vii) recommendations; and (viii) gender and human rights, including child rights (Annex 2 of the Global Evaluation Reports Oversight System, December 2012).

Using this scorecard, the review process generated three key data sets for each report: report typology, ratings (quantitative in nature), and comments/feedback on reporting quality (qualitative in nature). These data sets formed the basis for systematic comparison across the reports submitted.

Quality Assurance Process

As in previous years, the GEROS process was entrusted to a dedicated team of reviewers and senior reviewers, a coordinator, and a project manager. The coordinator and project manager worked closely with a small number of reviewers and UNICEF to oversee the process, ensure systematic completion of reviews, and respond to client and reviewer questions and requests in a timely manner. As with the 2013 cycle, we worked with a small team of reviewers. While this greatly facilitated the process of providing quality assurance, the team did require more calendar time to complete the reviews. As in the past, however, different members' skills and expertise (e.g. evaluation experience, language skills, knowledge of UNICEF) were leveraged with a view to ensure accuracy, depth, and consistency in both language and content. We sought to ensure consistency in ratings through three key steps:

- **Team briefings and discussions:** The review team met at the start of the process in order to become familiar with the review template, requirements, and expectations of the process. Subsequent discussions served to standardise ratings, as well as iron out any uncertainties regarding special or unique cases.
- **Senior and Peer Reviews:** Before reviews were submitted to UNICEF, the coordinator went through each one to ensure that all aspects had been completed. Following this, 20% of reports went through a peer review process, in which senior peer reviewers verified the depth, coherence, and quality of reviews according to the established expectations and objectives of the GEROS process. This proportion corresponds to that indicated in the GEROS guidance document.
  Special reports or reports flagged by UNICEF were reviewed by the more senior team members. In reviewing the work of their peers, team members had the opportunity to discuss and challenge ratings, ensure uniformity in the application of the template, correct inconsistencies and request further justification and supporting examples, as well as make any linguistic improvements needed.
- **UNICEF Feedback:** The UNICEF Evaluation Office was an inherent part of the quality assurance process, verifying each review to ensure consistency in the ratings and the quality of the narrative content/language. In a few cases, reviews were returned so that the team could further elaborate or support ratings and explanations.

4.2 Meta-Evaluation

A total of 69 reports were submitted to the UNICEF database for inclusion in the 2014 GEROS process. The data generated from the review of these reports forms the basis of the present meta-evaluation. The findings herewith were derived from the following process:

- **Data Aggregation:** As soon as UNICEF approved all of the reviews submitted, Universalia used a Visual Basic for Applications code in Excel to aggregate the ratings and comments into a single Excel workbook. The resulting document constituted the fundamental data set allowing for trends, strengths and weaknesses to be highlighted and compared.

- **Year-to-Year Comparisons:** To ensure consistency and coherence, the approach taken to analysing data and trends mirrored that of previous years. According to this approach, overall and section ratings were disaggregated based on report typology (region, geographic scope, management, purpose, result-level, level of independence, approach, Medium-term Strategic Plan (MTSP) (and as of 2014, SPOA) correspondence, and language). Resulting trends were analysed according to report typology, and the sections were categorised as per the four-point performance scale presented in Exhibit 4.1. It is worth noting that, in many cases, ratings of "Outstanding" and "Highly Satisfactory" were classified together in order to reflect all of the good quality reports that respect UNICEF standards. Expressed as percentage ratio, ratings for the different points on the scale were compared to the data obtained from 2010 to 2014. The graphs included herewith serve to illustrate year-to-year trends, notably the progress made as well as areas still requiring improvement.
- **Trends in the Overarching Sections:** Finally, both quantitative and qualitative data served to highlight trends within the overarching sections of the template (e.g. Section D – Findings and Conclusions, Section E – Recommendations and Lessons Learned, etc.). The ratings attributed to each sub-question of the template were compiled and disaggregated according to their sub-section (e.g. Completeness and Insight of Conclusions, Relevance and Clarity of Recommendations). The ratings per sub-section were then analysed in order to extract key strengths and weaknesses across the population of reports. Qualitative data was grouped according to key term or focus area, which helped pinpoint themes and support trends noted in the comparison of quantitative data.

4.3 Changes Made From Previous Years

Early 2010 marked the start of the GEROS process, in which reviews of 2009 reports were conducted. While the process remains very similar to this day, some changes were made following a rapid review by UNICEF in 2012, based on the experience and lessons of previous years. These modifications included new wording (but the same colour coding) in the four-point performance scale; new options ("Externally Managed" and "UNDAF Evaluation") under "Management of the Evaluation"; the addition of "Regional/Multi-Country Programme Evaluation" under "Purpose"; as well as the option to select both "Summative and formative" under the evaluation's approach (formerly called "stage").

For the 2014 cycle, very minor changes were made to the GEROS template. In the classification section of the tool, "MTSP Correspondence" and its categories were changed to "SPOA Correspondence" and its categories, in order to align with UNICEF's new strategic plan for 2014-2017.

4.4 Limitations

The GEROS team met with certain methodological and analytical challenges in reviewing reports and producing the present document. Although there are objective standards in the GEROS template, it is always necessary to use best judgment when applying those standards to each report, thus creating personal biases or expectations that may influence the results. To harmonize ratings across reviewers, previous lessons were shared internally and integrated into the process from the outset and peer reviews were conducted. The review team jointly reviewed a report at the beginning of the process to ensure that everyone had the same understanding of the standards. All GEROS templates were reviewed by the coordinator and UNICEF for quality control. The review team naturally came across unique or special cases that required deeper analysis or discussion in order to ensure consistency of ratings, such as impact evaluations, evaluation of training, and complex evaluations.

Indeed, unique or special cases – when they concerned an entire report, rather than just one or two template questions – necessarily posed challenges to the review team. While all the reports reviewed were considered evaluations, different evaluation types and designs put the flexibility of the template to the test, and may have also influenced rating consistency (see Appendix XIII for further detail). The team worked together to determine the most suitable ratings to attribute, but the experience prompted discussions around the applicability of the template and the different options that could be envisaged to increase its flexibility.

In some cases (although much improved from previous years), the evaluation TORs were not included to help complement or justify certain choices, approaches or foci in the reports (e.g. those that differed from expected standards or traditional practice). While GEROS requirements may have been satisfied in other deliverables, the exclusion of these deliverables from the final report may have affected the ratings assigned.

Finally, as noted in previous years, changes made to the template – such as the wording of the colour-coded scale in 2012 - may have also influenced the ratings attributed (due to differences in interpretation) and thus, the ability to compare results over multiple years. When comparisons across years were not possible, the report explicitly describes why that is the case.

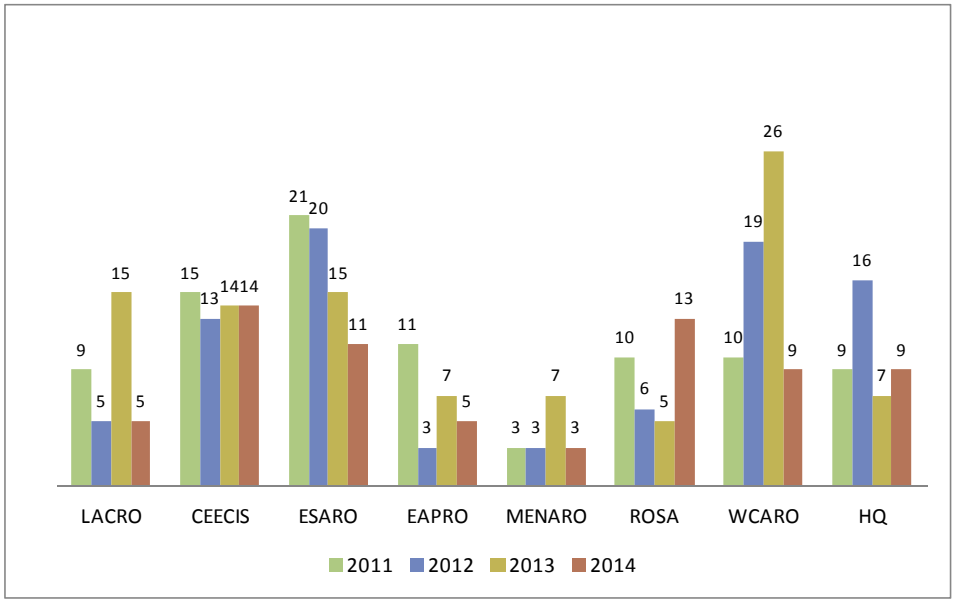
5 Findings

Based on the analysis of the data obtained from the review process, the following findings emerge.

5.1 Overall Ratings and Feedback

Between the 2013 and 2014 GEROS exercises, the number of reports reviewed decreased from 96 to 69 reports. As demonstrated by Exhibit 5.1, the number of reports submitted by each region has varied – in some cases, quite significantly – over the last four GEROS cycles.[5] While the number of reports from the Central and Eastern Europe/Commonwealth of Independent States (CEE/CIS) Regional Office has stayed relatively stable and now leads in terms of the number of reports submitted, reports submitted from the East and Southern Africa Regional Office (ESARO) has continued to decline. The number of reports from the Regional Office of South Asia (ROSA) and HQ increased somewhat compared to last year. This total number of report has to be understood in relative terms. Three multi-country/regional evaluations are included in that sample, which may increase country coverage in some regions. Two multi-country evaluations were completed by CEE/CIS, while one regional evaluation was carried out by the Latin America and Caribbean Regional Office (LACRO). Multi-country evaluations by CEE/CIS cover about five countries, which further increases the coverage for that region.

Exhibit 5.1 Reports Reviewed per Region per Year



[5] We understand that additional evaluation reports were submitted after the cut-off date (May 1st, 2015) and thus are not reflected in this aggregate analysis. In addition, other types of evaluative reports produced by country offices are not reflected in this analysis, such as research, reviews, and assessments, which may provide feedback on performance, but are not considered evaluations as per UNICEF's definition.

Finding 1: In 2014, the quality of reports submitted continued to increase overall, accompanied by a large increase in the number of reports with TORs included in the appendices or as separate documents.

For the purposes of the GEROS meta-analyses, good quality reports are those rated as highly satisfactory or outstanding. As demonstrated in Exhibit 5.2 below, the number of good quality reports has grown consistently over the years, and continued to do so in 2014, with the largest proportion of good quality reports (74%) since the GEROS exercise began. This suggests a generalised improvement in reporting quality. However, it should also be noted that, this year, regions with declining quality of evaluation reports also contributed many fewer reports to GEROS this year (with the exception of LACRO), making up a smaller proportion of the report sample.

Exhibit 5.2 Good Quality Evaluation Reports over Time (2009-2014)

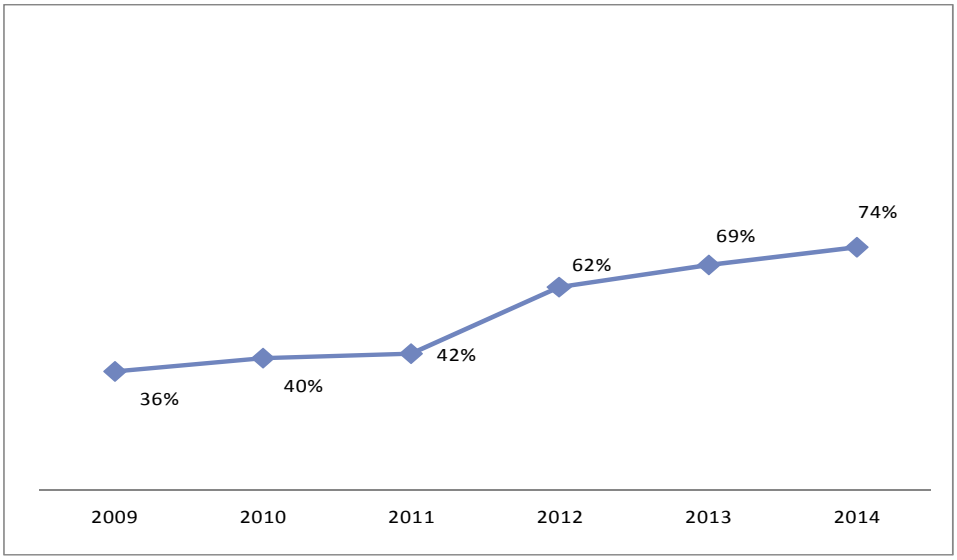
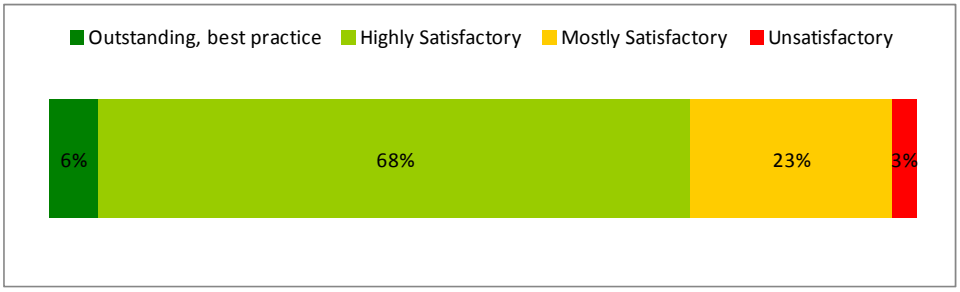


Exhibit 5.3 shows this moderate overall improvement of report quality: highly satisfactory reports increased from 64% last year to 68%, and mostly satisfactory reports decreased from 29% to 23%. Outstanding and unsatisfactory reports remained at approximately the same level.

Exhibit 5.3 Overall Ratings for 2014



Unsatisfactory and mostly satisfactory ratings indicate the absence or poor execution of GEROS standards. Unsatisfactory was used when a standard was poorly addressed, off track or completely absent. Mostly satisfactory was used when a standard was partially addressed.

Prior to 2014, TORs were often missing, which made cross-verification of reporting requirements difficult. However, in 2014 there was a large improvement in the number of TORs present, with 88% of reports being accompanied by their TORs, compared to only 57% in 2013. The inclusion of the TORs allows the reviewers to more accurately assign ratings to a report, and may contribute to an increase in ratings, as in previous years reports with TORs tend to be rated more highly than reports that did not.

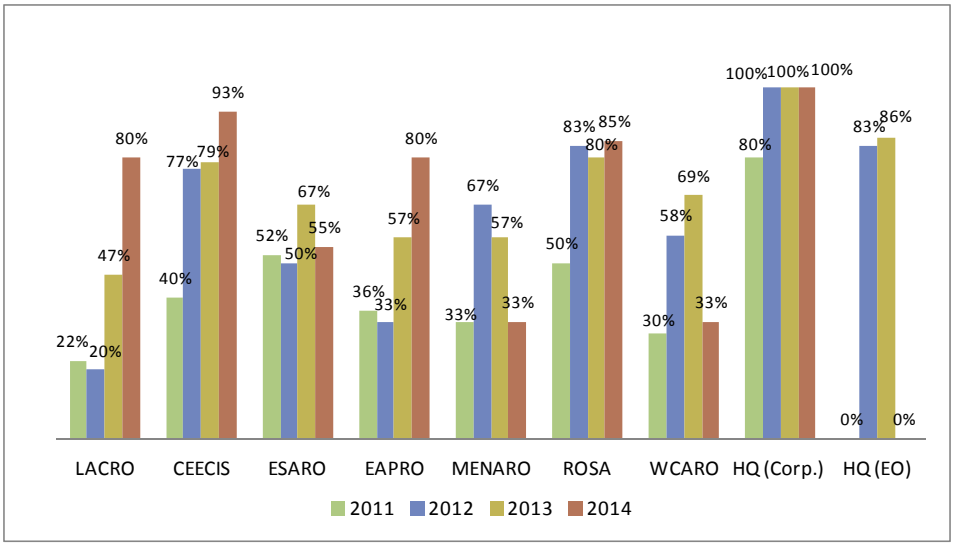
5.2 Overall Regional Trends

Finding 2: The quality of UNICEF evaluation reports varied since last year's review process, with some regions improving and others declining.

Data analysis in 2014 indicates that regions leading in report quality include HQ (Corporate), CEE-CIS, and ROSA. There are varying patterns between the regions in the quality of reports submitted over time, with some regions improving, some losing gains from previous years, and others remaining the same (Exhibit 5.4). HQ Corporate continued to have universal good quality ratings, and ROSA maintained a similar level of good quality reports to that of the past. CEE-CIS had a jump in good quality reports and now reaches 93% of good quality reports. The West and Central Africa Regional Office (WCARO), and ESARO had noticeable declines. The East Asia and Pacific Regional Office (EAPRO), the Middle East and North Africa Regional Office (MENARO) and LACRO submitted too few reports to draw firm conclusions.

Closer observation of the overall ratings attributed within each region, shows that CEE-CIS had the highest number (3) of outstanding reports (in fact, most regions did not have any reports with overall outstanding ratings). The proportion of highly satisfactory to mostly satisfactory reports varied between regions, but only WCARO and ROSA registered any unsatisfactory reports (one each) in 2014.

Exhibit 5.4 Proportion of Good Quality Reports per Region[6] [7]



An important observation of the linkages between quality and quantity of regional reports in this year's sample is that most of the regions with declining quality of evaluation reports (WCARO, ESARO, and MENARO) also contributed many fewer reports to GEROS this year, making up a smaller proportion of the report sample. This raises the question of whether there is any relationship between good quality ratings and the contribution of reports to the GEROS data base. There are a number of factors that may explain varying levels of contribution of reports by region. In some cases, fewer evaluations may be conducted in one year compared to another. This is sometimes related to a low implementation rate in a country where a project or programme has not progressed enough to justify an evaluation. Evaluations may be planned but not all conducted due to issues with budgets, lack of priority, or staff turnover. Finally, in some cases low quality evaluations may be reclassified as a review and subsequently not submitted to GEROS.

5.3 Trends by Type and Scope of Evaluation

Finding 3: As in previous years, most reports focused on initiatives with a national scope and they were overall good quality reports.

Geographic scope refers to the area covered by the programme or project being evaluated, and helps determine the extent to which evaluation findings can be generalised. Once again, the great majority of reports reviewed covered a national scope (88%)[8]. Quality of national reports reached 72%.

Reports with a multi-region/global scope[9] are the next largest category of reports, at only 7% of the sample. All achieved a good quality rating, which matches the 100% good quality ratings from 2013

[6] The percentage of good quality reports was calculated by adding the number or reports that were outstanding and highly satisfactory per region over the total number of reports per region. Total number of reports per region can be seen in Exhibit 5.1.

[7] HQ Corporate reports refer to evaluations that are commissioned and managed from Headquarters, but not by the Evaluation Office. HQ Evaluation Office reports are those that are commissioned and managed by the Evaluation Office. In 2015, these two categories have been amalgamated.

[8] Includes reports that were classified as "national" and "sub-national".

for this category. The presence of only one report with a regional[10] scope rated as "mostly satisfactory" does not allow for meaningful comparisons. The two reports with a multi-country[11] scope in 2014 remained good quality, as was the case for the single report of this type in 2013.

Finding 4: Most evaluation reports in the sample were managed by UNICEF and were considered quality reports. With respect to evaluation purpose, programme and project evaluations continue to represent the most important proportion of evaluations reviewed.

In the majority of reports reviewed, UNICEF is the manager of the evaluation and has direct responsibility for the evaluation process; in others, UNICEF jointly manages the exercise with countries and/or with other agencies. In 2014, the proportion of UNICEF-managed reports[12] continued to increase, from almost half of the reports in 2013 to 80% of reports in 2014. These UNICEF-led reports also increased in quality, from 74% to 80%. This high level of quality might reflect the greater level of control that UNICEF staff can exercise over the adherence to standards and practices in evaluations that are managed by them.

In this cycle, there are 14 evaluations that were jointly managed[13]. A majority of joint evaluations with countries were considered to be of good quality (63%). The absence of country-led evaluations is notable given the importance given to country-led evaluation in the 2013 UNICEF Evaluation Policy[14].

Just as in last year's cycle, the 2014 GEROS exercise noted a large proportion of programme (32%) and project (28%) evaluations among the reports reviewed. These proportions represent a decrease compare to last year's. The proportion of pilot, project, and policy evaluations increased noticeably, while at-scale and impact evaluations increased slightly. The proportion of humanitarian evaluations decreased slightly, and this small group included one real-time evaluation (which was found to be highly satisfactory overall).

In 2014, project-level evaluations increased significantly from 58% good quality ratings to 84%. Programme-level evaluations also increased in quality, but by a smaller 7 percentage points, with a 70% level of good quality ratings. Reports in a few other categories struggled to meet GEROS standards, including the small groups of at-scale and impact evaluations[15] (see Exhibit 5.5).

---

[9] The programme is implemented in two or more regions, or deliberately targets all regions. The evaluation would typically sample several countries across multiple regions, with the results intended to be generalizable in two or more regions.

[10] Where one programme is implemented in several countries, or different programmes of a similar theme are implemented in several countries, the evaluation covers multiple countries within the region and the sampling is adequate to make the results generalizable to the region.

[11] The programme is implemented in two or more regions, or deliberately targets all regions. The evaluation would typically sample several countries across multiple regions, with the results intended to be generalizable in two or more regions.

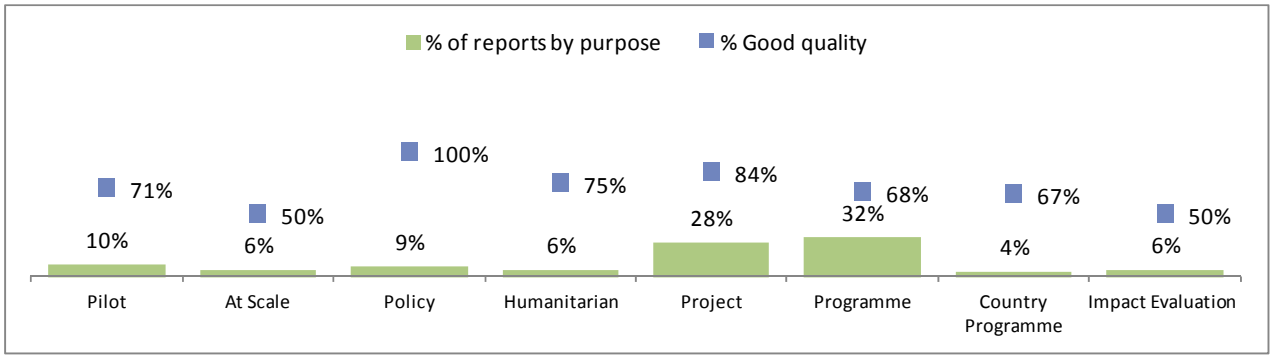
[12] UNICEF-managed evaluations are those in which UNICEF, working with national partners of different categories, is responsible for all aspects of the evaluation.

[13] Joint-evaluations can be: 1) managed with one or more UN agencies; 2) managed with other organizations or 3) can jointly managed by the Country (Government and/or CSO) and the UNICEF CO.

[14] Country-led evaluation: Evaluations managed by the Country (Government and/or CSO).

[15] A contributing factor to performance may be related to the different expectations for these types of evaluations, which are not necessarily reflected in the GEROS framework.

Exhibit 5.5 Proportion of Reports Reviewed and Good Quality Reports by Purpose (2014)[16]



Finding 5: A larger proportion of evaluations are summative this year and a majority of reports focus on education as a thematic area.

In 2014, the proportion of reports by type of evaluation approach differed somewhat from the previous two years. While the proportion of formative evaluations remained almost the same, summative evaluations increased this year by 9 percentage points, and evaluations combining a formative and summative approach[17] decreased by the same proportion. The quality of formative evaluations has improved tremendously, from 64% in 2013 to 85% in 2014, while quality of summative evaluations has slightly decreased. Evaluations combining a formative and summative approach, increased slightly in quality. [18]

In addition to the approach, the GEROS template asks reviewers to identify which thematic areas are addressed by the object of the evaluation. Prior to 2014, the template used the focus areas from the Medium-term Strategic Plan (2006-2013). In 2014, the Strategic Plan Objective Area (SPOA) list of focus area priorities was adopted, reflecting the new 2014-2017 strategic plan thematic priorities, which are somewhat different. Thus comparisons of trends over previous years are limited, as the table below shows:

Exhibit 5.6 Changes in MTSP and SPOA Correspondence categories

MTSP Correspondence (2013)	SPOA Correspondence (2014)
HIV/AIDS and children	HIV/AIDS
Child Protection	Child Protection
Basic education and gender equality	Education
Young child survival and development	Health
Multi-sectoral	Nutrition

[16] Total number of reports per purpose: 7 pilot, 4 at scale, 6 policy, 0 RTE, 4 humanitarian, 19 project, 23 programme, 3 country-programme and 4 impact evaluations. The definition of each of these terms can be found in Appendix XII (GEROS criteria).

[17] The category "formative and summative" evaluation was only added to the GEROS process in the 2012 exercise in order to fully reflect the breath of reports reviewed.

[18] Total number of reports for each approach: 20 formative, 29 summative, 20 formative and summative.

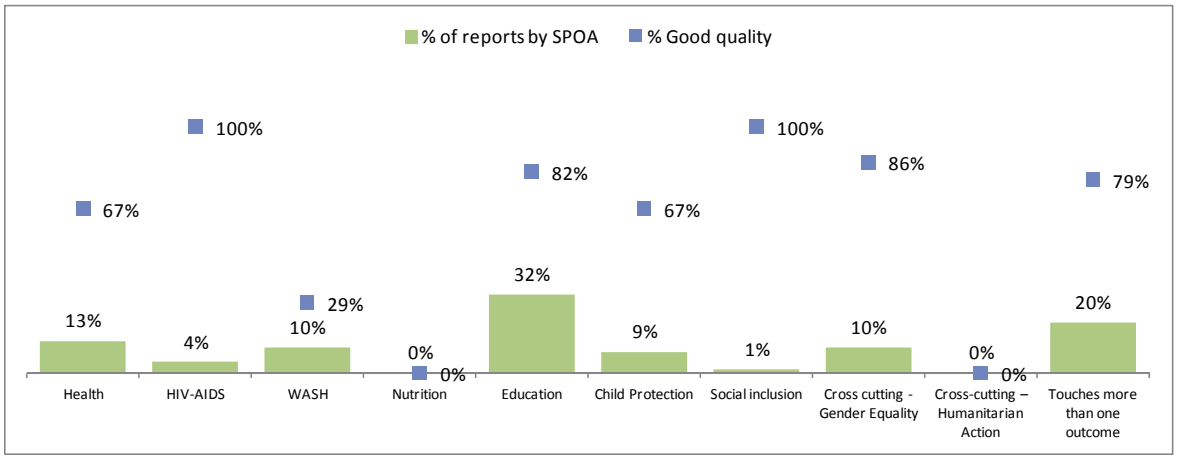
MTSP Correspondence (2013)	SPOA Correspondence (2014)
Cross-cutting	WASH
Policy Advocacy and partnerships	Social inclusion
Organizational performance	Cross-cutting – gender equality
	Cross-cutting – humanitarian action

Themes which have remained the same are HIV/AIDS, child protection, and multi-sectoral (evaluations that touch on more than one thematic outcome area). Of these, the number of HIV/AIDS and child protection thematic evaluations are very similar in proportion to the 2013 sample. Evaluations that touch on more than one thematic outcome area have dropped 8 percentage points from 28% to 20%.

Education is the most common theme addressed in evaluations in 2014, making up 32% of GEROS reports (Exhibit 5.7). More modest but still significant numbers of reports in 2014 were reviewed in health (13%), WASH (10%), and cross cutting-gender equality (10%), in addition to child protection as noted above (9%).

These numbers raise some questions regarding the extent of coverage of provided to HIV/AIDS, nutrition, social inclusion, and cross-cutting humanitarian action, which are key focus areas of UNICEF in the 2014-2017 Strategic Plan (some of which were combined with other categories in previous years). It may be worth studying whether the low number of reports submitted on these topics reflects an evaluation coverage issue, a limited number of UNICEF initiatives in these areas, or a natural fluctuation over time (for example, nutrition was a particular focus in 2013 in a number of regional and global studies). UNICEF may also wish to consider whether quality issues result from a lack of resources for Monitoring and Evaluation (M&E) of its WASH work.

Exhibit 5.7 Proportion of Reports Reviewed and Good Quality Reports by SPOA Correspondence[19] [20]



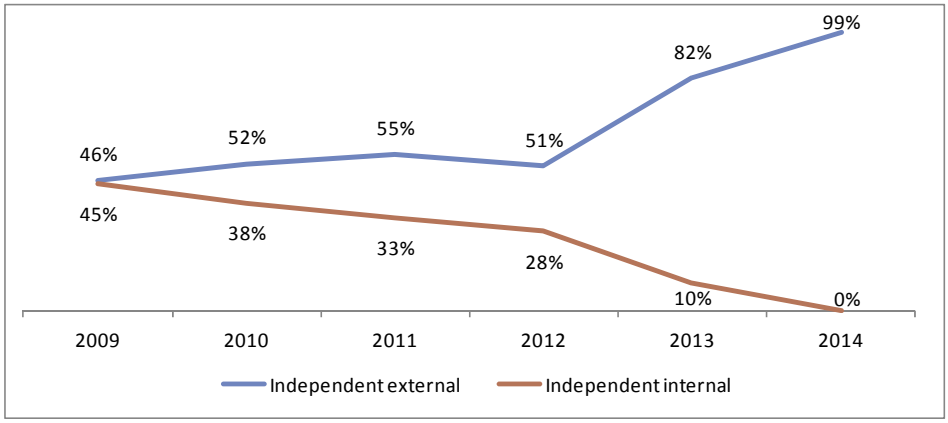
[19] Total reports per SPOA Correspondence: 9 health, 3 HIV/AIDS, 7 WASH, 0 nutrition, 22 education, 4 child protection, 1 social inclusion, 6 cross-cutting gender equality, 0 cross-cutting humanitarian action, 11 evaluations with more than one outcome.

[20] There is no evaluation that has a SPOA correspondence that relates to humanitarian action. The four reports that had a humanitarian purpose had different SPOA correspondence (focus area priorities). These reports were addressing 1) gender equality and 2) WASH, in emergency settings.

Finding 6: Almost all reports submitted to GEROS are independent external evaluations, which is a shift since 2012 when internal evaluations (or reviews) were still sometimes classified as evaluations. A majority of the reports submitted were in English.

The level of independence refers to where the implementation and control of the evaluation activities are held. A steady decline of independent internal reports[21] has eliminated this type of report from the GEROS sample (Exhibit 5.8). The sample contains independent external evaluations, except for one report, for which the level of independence could not be determined (compared to 22% of reports in 2013 for which this could not be determined). Part of this shift likely reflects that a growing number of reports clearly identify the evaluators and the office responsible for managing the evaluation. It may also be indicative that internal evaluations are more frequently considered 'reviews' and are therefore not submitted to GEROS.

Exhibit 5.8 Proportion of Reports by Level of Independence (2009-2014)



Furthermore, English remains the primary reporting language in the GEROS process at 89% of reports reviewed. The number of French and Spanish reports declined significantly from 2013, at 9% and 3% respectively[22]. This drop in language diversity may reflect the fewer number of reports submitted by regions such as WCARO and LACRO.

This small pool of non-English reports makes quality comparisons less relevant. However, we can make the observation that the two Spanish reports were both found to be highly satisfactory, while the six French reports declined somewhat in good quality ratings.

[21] As per the GEROS template definitions of these terms (Appendix XI), an independent internal evaluation is one that "is implemented by consultants but managed in-house by UNICEF professionals. The overall responsibility for the evaluation lies within the division whose work is being evaluated." An independent external evaluation is "implemented by external consultants and/or UNICEF Evaluation Office professionals. The overall responsibility for the evaluation lies outside the division whose work is being evaluated."

[22] Total number of reports per language: 61 English, 6 French and 2 Spanish.

5.4 Trends by Quality Assessment Category

5.4.1 Overall Trends All Sections

The GEROS template is divided into six subsections, each on a particular aspect of UNICEF reporting standards:

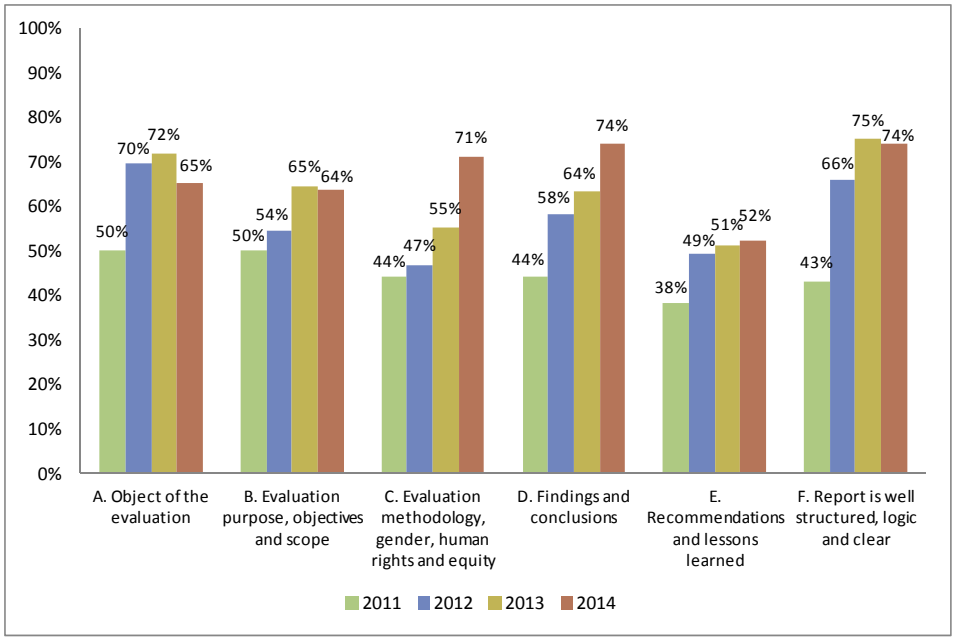
- Section A comprises the description of the evaluated object;
- Section B relates to the report's purpose, objectives, scope and evaluation framework;
- Section C covers methodology and inclusion of human rights, gender and equity;
- Section D includes findings and conclusions,
- Section E assesses recommendations and lessons learned; and
- Section F addresses the report's structure, layout and logic.

Each of these sections is made up of varying numbers of individual questions organized into sub-sections (see GEROS template in Appendix VI). The following pages analyse template subsections A through F, and their sub-section ratings, in order to underscore specific areas of strength and weakness to be maintained or addressed in the coming years.

Finding 7: In 2014, the quality of reports slightly increased in two sections, decreased slightly in another, and changed little in three other sections. The majority of reports continue to be aligned with the UNICEF-adapted UNEG standards for evaluation reports.

Exhibit 5.9 illustrates that, until 2014, the different report sections demonstrated continuous improvement, progressively achieving greater alignment with UNICEF reporting standards. Sections C (methodology, gender, human rights, and equity) and Sections D (findings and conclusions), which have improved since last year, are some of the highest-rated standards in the template, despite being amongst the lowest-rated in the early years of the GEROS exercise. Section A (Object) was found to be slightly less well done than in 2013. The other sections remain largely unchanged in terms of the proportion of reports achieving an outstanding or highly satisfactory rating. For their part, recommendations and lessons learned (Section E) have consistently been – and still remain – among the most challenging report standards to meet.

Exhibit 5.9 Good Quality Ratings per Section – Year by Year Progression (2011-2014)

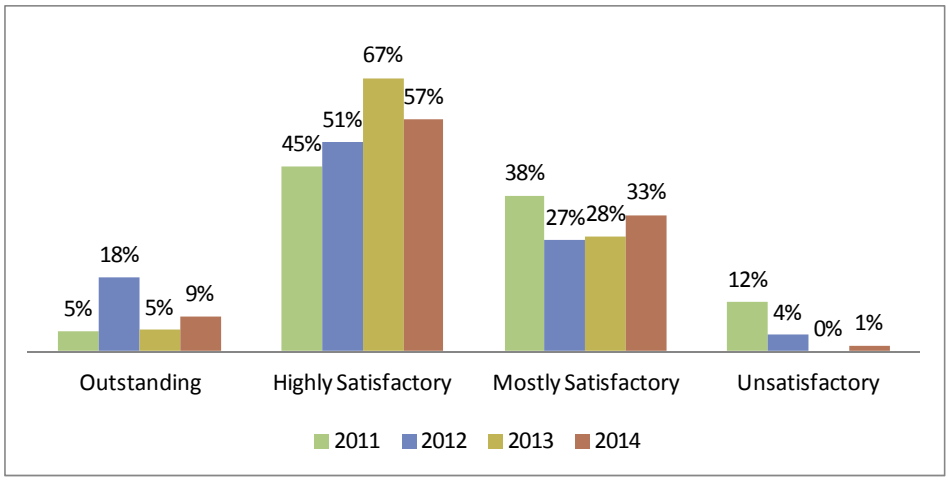


5.4.2 Section A: Object of the Evaluation

Finding 8: In 2014, the description of the evaluated object and its context declined somewhat in quality compared to 2013. Evidence suggests that the description of the theory of change and of stakeholder roles and contributions remain areas for improvement.

The proportion of reports providing a good quality description of the evaluated object and its context decreased somewhat in 2014 (from 72% to 65%), which is considerably better than the 2010-2011 ratings for this section (Exhibit 5.10).

Exhibit 5.10 Section A Ratings and Comparison (2011-2014)



The sub-section ratings within this section (Exhibit 5.11) reflect an increase in the description of the theory of change, decrease in describing the object and context, and relatively unchanged ratings in outlining stakeholders and their contributions compared with last year.

It shows that a majority of reports (73%) continued to do well (in other words, to be rated as outstanding or highly satisfactory) in terms of presenting a good description of the **object and context.** Reviewer comments in the template noted a good description of context , and provided compliments for adequately outlining the components of the evaluated object. These reports were often commended for providing the institutional, social, political, economic, demographic, and policy context related to the object of the evaluation, which provided important context for the evaluation findings.

Example of review of an Outstanding Section A

"The report provided a very clear and thorough description of the object. This description included a full explanation of the project's theory of change and provided the reader with a good understanding of exactly how the Project was designed to address the problem described in the context. The context was also fully described, with information provided about the political, economic, demographic, and specific child protection policy and legal frameworks. Key stakeholders and their contributions were clearly identified and described."

Final Evaluation of the Project "Child Care System Reform" (GEROS-Montenegro-2014-004)

Example of review of a Mostly Satisfactory Section A

"L'objet de l'évaluation est relativement bien décrit mais reste incomplet. La chaîne des résultats attendus est manquante ainsi que le rôle précis des divers intervenants. Les parties prenantes du projet et de l'évaluation sont mentionnées mais l'information sur les contributions précises des intervenants clés est très limitée."

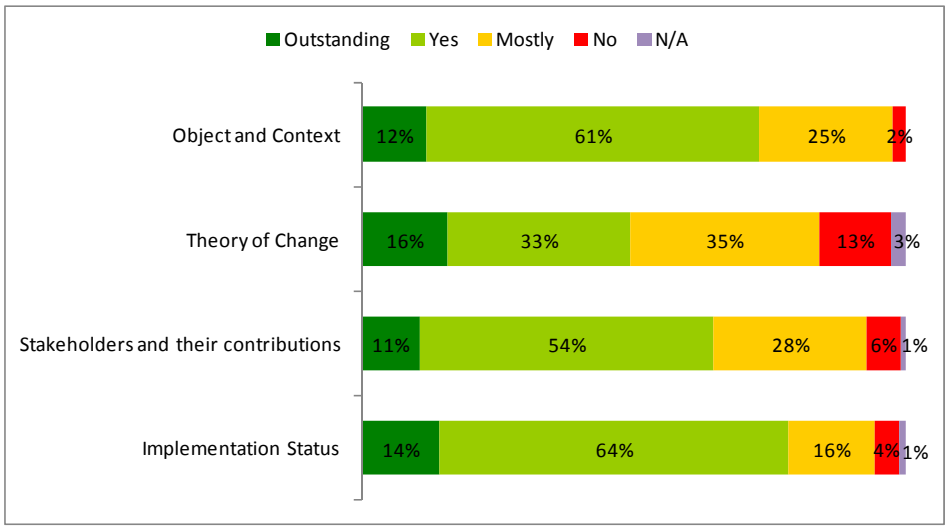
"Evaluation du programme élargi de vaccination dans les camps de réfugiés sahraouis de Tindouf" (GEROS-Algeria-2014-003)

However, the description of the theory of change and the presentation of stakeholder contributions remain areas for improvement. The proportion of reports with good quality ratings for the sub-section of **stakeholders and their contributions** (65%) remained essentially unchanged from last year. However, this amalgamated rating for the sub-section masks the large difference between the two standards making up this sub-section: identification of stakeholders in the report (which is done well, with 86% being good quality) and the description of key stakeholders contributions (only 54% good quality). Indeed, almost one-third of reviewer comments noted that contributions of stakeholders were not clear.

The sub-section which continues to be the weakest, relating to the **theory of change**, nonetheless showed improvement this year, rising 8 points to 49%. One-third of reviewer comments noted that the theory of change or results chain was not made clear in the report. In some cases, this was because the evaluated object did not have one, but this was overcome in some cases by evaluation teams who developed a theory of change, and used this to help guide the evaluation.

Finally, a majority of reports (78%) did well in describing the **implementation status** of the evaluated object. Reviewers noted however some instances in which significant changes or modifications to the programme were not thoroughly described in the evaluation report.

Exhibit 5.11 Sub-Section Ratings: Object of the Evaluation

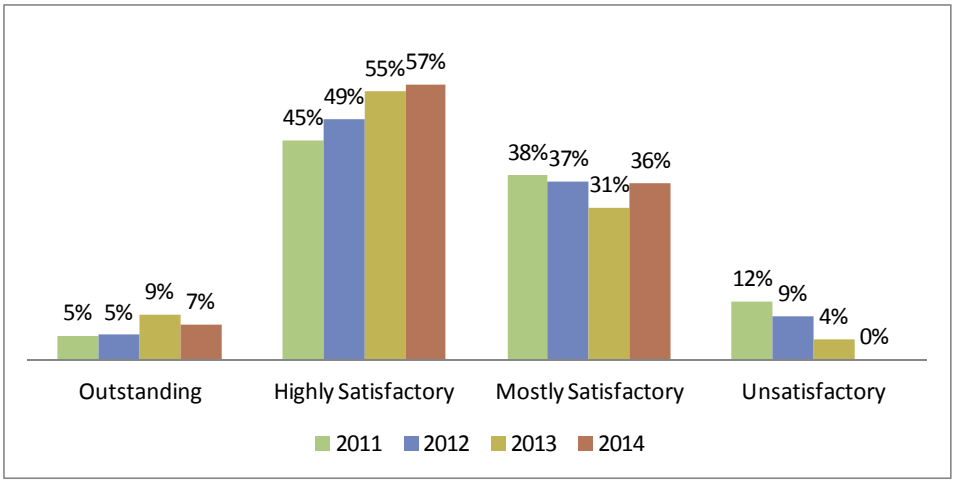


5.4.3 Section B: Evaluation Purpose, Objectives and Scope

Finding 9: The extent to which reports met standards of evaluation purpose, objectives and scope changed little from previous years. Strengths of reports in this area included clear purpose, objectives and scope and a clear list of evaluation criteria. However, the justification for the selection of evaluation criteria still requires greater attention.

Section B of the GEROS template deals with the evaluation's purpose, scope, and objectives, as well as the presentation of the framework guiding the assessment. Although this is not amongst the strongest sections, most reports (64%) are considered to be good quality in terms of these standards. The quality of Section B changed little from previous years. Perhaps the most notable trend is the slow but steady drop (to zero, in 2014) in the number of unsatisfactory reports in this section.

Exhibit 5.12 Section B Ratings and Comparison (2011-2014)



The distribution of individual sub-section ratings (Exhibit 5.13) has not changed significantly compared to 2013. The description of the evaluations' **purpose, objectives and scope** remained good quality in three-quarters (76%) of reports. Indeed, reviewer comments commended reports for their description of the evaluation's purpose, and for clearly explaining objectives. Good quality reports tended to link these clearly together. In some cases, weaker reports did not address these explicitly, particularly in the case of scope.

Example of review of an Outstanding Section B

"The evaluation's purpose, objectives, and scope, including all evaluation criteria used in the context of this evaluation, are very clearly explained and presented. Differences between the initial TOR and what was agreed to at inception are fully explained. The report provides a relevant list of key evaluation questions."

Evaluation of "Young Champions Initiative for Girls' Education" (GEROS-Pakistan-2014-002)

Example of review of a Mostly Satisfactory Section B

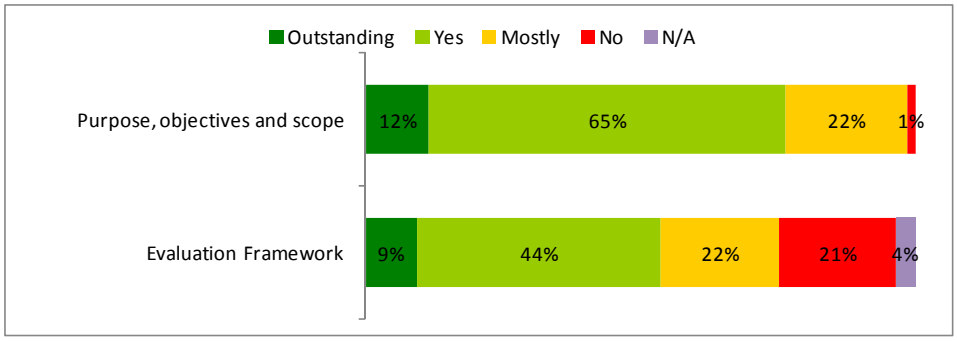
"The report suggests but does not clearly describe the purpose, objectives, and scope of this evaluation. The reader is left to glean this information from different parts of the report. Moreover, the report does not provide an explanation as to why the standard OECD-DAC evaluation criteria were not used in the context of this evaluation."

"Impact evaluation of the WASH SHEWA-B programme in Bangladesh" (GEROS-Bangladesh-2014-014)

The **evaluation framework** subsection also changed little from last year, and continues to be amongst the weaker sub-sections in the template with only 54% of reports rated good quality for this standard. However, the weakness in this sub-section is not the listing of the evaluation criteria themselves (which 69% of reports did well), but the frequent lack of justification for why particular evaluation criteria were chosen. An equal number of comments were made noting that the criteria were, or were not, explicitly justified.[23]

[23] An amendment was made to the wording of the template in 2013 to recognise that the use of standard OECD/DAC criteria required less justification than the use of other criteria. However, some reviewer comments indicate that the extent to which some justification of even OECD/DAC criteria should be provided is not clear, and the ratings may be inconsistent in this regard.

Exhibit 5.13 Sub-Section Ratings: Purpose, Objectives and Scope of the Evaluation

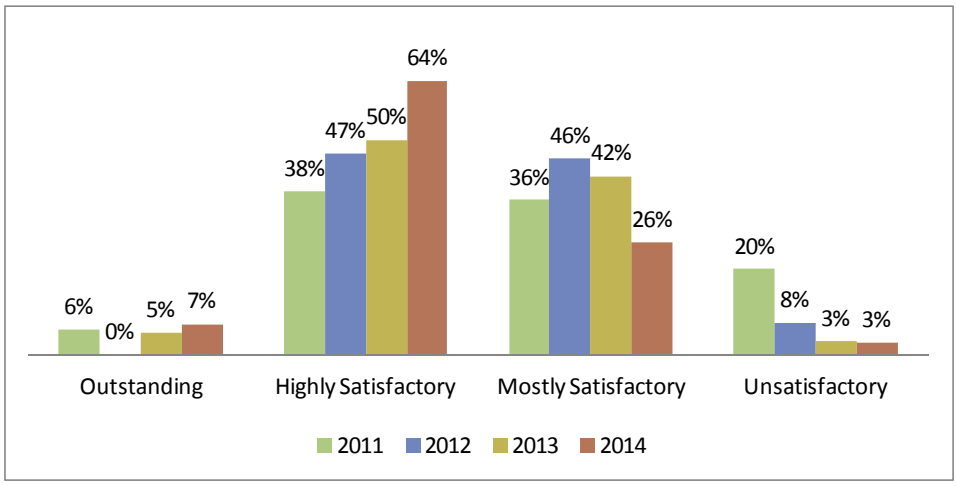


5.4.4 Section C: Evaluation Methodology, Gender, Human Rights and Equity

Finding 10: Reports have made improvements in terms of the description of the methodology. While methodological robustness is often satisfactory, lacunas in ethical considerations and stakeholder participation may have impacted the overall ratings of this section.

Section C, on report methodology and integration of human rights, gender and equity, has improved over time, but particularly so in 2014. A much larger proportion of reports were considered to be good quality in this section (71% compared to 55% in 2013).

Exhibit 5.14 Section C Ratings and Comparison (2011-2014)



Looking more closely at the sub-sections (Exhibit 5.15), increases in the number of good quality ratings are noticeable. With the exception of data collection, good quality ratings for all sub-sections increased between 2013 and 2014. This year, ratings seem to be converging, with four out of six sub-sections receiving a similar proportion of good quality ratings (between 66% and 72%). The 24-point improvement in the stakeholder participation ratings is the most notable.

Example of review of an Outstanding Section C

"The report offers an excellent description of the methodology and the annexes serve to complement this description. Specifically, the data collection, analysis and sampling methods are described and a justification for using these is given. The evaluation matrix helps the user understand how the evaluation was designed to answer specific criteria and questions and ultimately achieve the evaluation objectives and purpose. Moreover, the construction of the TOC was used to guide the entire evaluation, and better assess effectiveness and other criteria."

Final Evaluation Report "RKLA3 Multi-country evaluation: increasing access and equity in early childhood education" (GEROS-Regional Baltic-2014-008)

Data collection, for which good quality ratings dropped 7 points from 2013 to 72%, is still the strongest sub-section. Every single report reviewed included at least some description of data collection, methods, and data sources. This was directly reflected in reviewer comments, which complimented reports on their clear identification or description of the data collection methods.

Example of review of a Mostly Satisfactory Section C

"La description sommaire de la méthodologie identifie des sources et méthodes de collecte et d'analyse des données. Toutefois, la méthodologie n'est pas pleinement décrite et les outils de collecte de données ne sont pas inclus en annexe du rapport."

Évaluation de "l'approche assainissement total piloté par la communauté" (GEROS-Madagascar-2014-005)

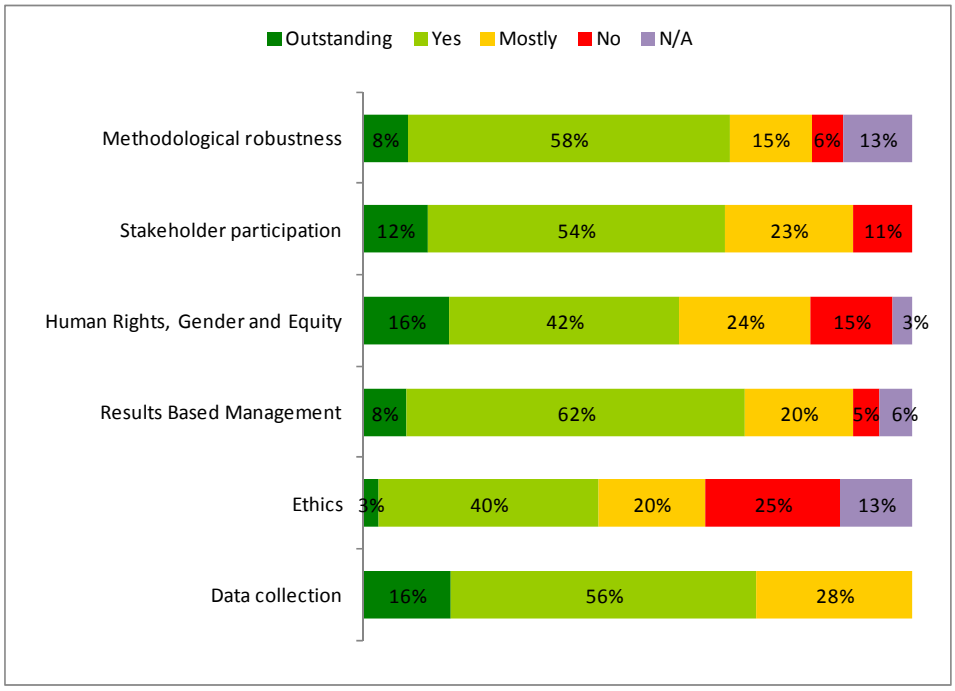
The sub-section addressing **results-based management** (RBM) was also an area of strength (with 70% of reports being aligned with UNICEF standards), and an improvement from last year by 10 points. Positive ratings indicate that reports did an adequate job of evaluating the object's monitoring system, and that the evaluation made appropriate use of the M&E framework. Often the M&E framework was used as a guide to the analysis and sometimes for the structure of the report.

Similar to the RBM sub-section, most reports (67%) received good quality ratings for **methodological robustness**, an increase of 5 points from 2013 for this sub-section. The individual standards making up this section show a strength in terms of having the methodology facilitate answers to the evaluation questions (81% having done this well). This sub-section would receive higher ratings but for the frequent absence of a counterfactual to address issues of contribution or attribution (only 47% having done this well). In many cases, reviewers commented that a counterfactual was impractical or not relevant, and 37% were rated Not Applicable. Certainly, this is an area for more specific guidance in TORs, a greater understanding for evaluators of what is entailed in constructing a counterfactual, and when it may or may not be appropriate.

Stakeholder participation improved significantly this year, rising 25 percentage points from 2013 to 66%. It is not clear what has driven this improvement overall. However, the individual standards that make up this section indicate that reviewers found in most cases levels of participation in the evaluation were appropriate (71% good quality ratings), while there were more often gaps in actually describing these levels (60% good quality ratings). Only one fifth of reviewer comments explicitly noted that stakeholder participation was well described. The higher ratings reflect that reviewers were often able to infer stakeholder participation from the report, even if it was not explicitly described.

The sub-sections that remain weak are a) ethics and b) human rights, gender and equity. The **ethics** sub-section demands a description of ethical considerations in the programme as well as in the evaluation process (e.g. protection of confidentiality, informed consent). The ethics sub-section received the smallest proportion of good quality ratings in this section, as well as the fewest number of outstanding ratings. One-quarter (25%) of reports neglected to discuss ethics at all (and thus received a rating of unsatisfactory). The inclusion of considerations of the ethical design of the programme, the balance of costs and benefits to participants, and the ethics of who was included and excluded in the evaluation was particularly poorly done, with only 43% of reports receiving a good quality rating for this individual criterion.

Exhibit 5.15 Sub-Section Ratings: Methodology, Gender, Human Rights, and Equity



Human Rights, Gender, and Equity

In its reports, UNICEF places considerable importance on the inclusion of human rights, gender and equity considerations, from the methodology, through to the recommendations. In 2012, the integration of a Human Rights-Based Approach to Programming (HRBAP), gender equality and equity considerations in the reports reviewed was still considered as a weakness, with none of the three elements receiving more than 46% good quality ratings. With the exception of gender, improvement was made in 2013 and 2014, and all three considerations either met or marginally surpassed the halfway mark (50%) in 2013 and 2014 (Exhibit 5.16).

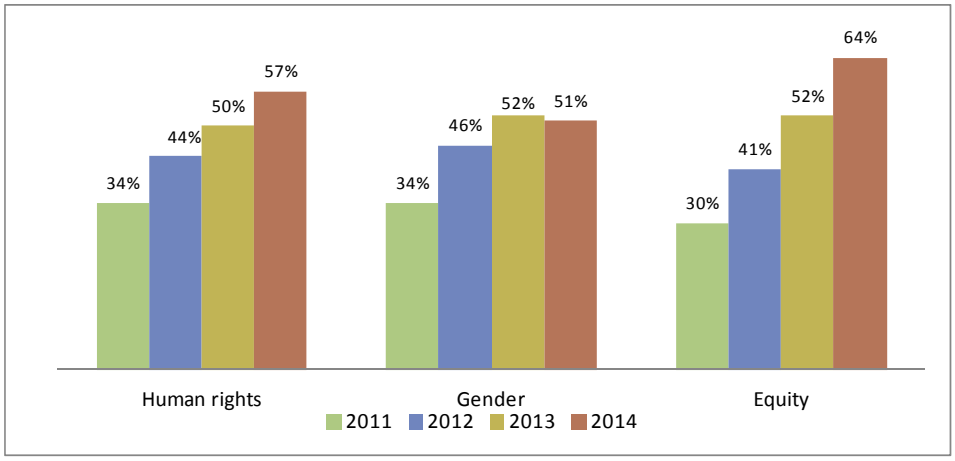
When considered as a whole, the sub-section pertaining to human rights, gender and equity considerations increased somewhat in quality from 2013. When some of the standards of this sub-section (pertaining specifically to inclusion of either human rights, gender, or equity in the methodology, framework, findings, conclusions and recommendations) are analysed individually, however, it is evident that the recent progress has been made in the inclusion of equity and human rights. This continues the pattern of ongoing improvement of inclusion of **equity** in reports, which climbed from 9% good quality ratings in 2010, to 64% in 2014. Equity now has the highest

proportion of good quality ratings of the three themes. As noted in last year's meta-analysis, such a drastic hike may be the result of UNICEF's efforts and investments around equity since 2011.

Good quality ratings for **human rights** also improved somewhat, from 50% in 2013 to 57% in 2014. Reviewer comments praised the inclusion of human rights, noting reports that cascaded human rights considerations throughout the report, or used human rights language or frameworks. In other reports, these approaches were absent, or their inclusion was weak or appeared as an afterthought in the report.

The inclusion of **gender** in the methodology, framework, findings, conclusions and recommendations did not change significantly compared to 2013. Partial points were given for reports that addressed gender in some ways such as including gender-disaggregated data, while stronger reports included, for example, a comprehensive section on gender considerations, or used gender equality as a guiding principle of the evaluation. Over the coming years, it will be interesting to note how the gender component improves in response to the UN System-Wide Action Plan on Gender Equality and Women's Empowerment, adopted in 2012. So far, continuing improvement on gender is not evident in reports assessed through GEROS.

Exhibit 5.16 Inclusion of Human Rights, Gender, and Equity: Good Quality Ratings Year by Year

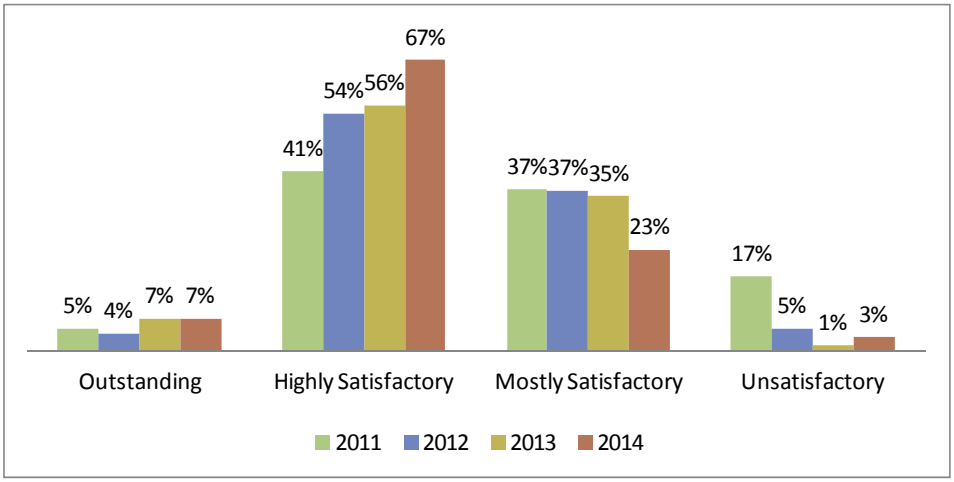


5.4.5 Section D: Findings and Conclusions

Finding 11: The presentation of findings, conclusions, contribution and causality continued to improve in 2014. Cost analysis remained particularly challenging.

In 2014, Section D increased in quality, with a greater proportion of reports considered good quality (from 64% in 2013 to 74% in 2014) (Exhibit 5.17).

Exhibit 5.17 Section D Ratings and Comparison (2011-2014)



In looking at the sub-section ratings, all but one (cost analysis) increased moderately in good quality ratings. Strengths, weaknesses and implications remained a particular strength in this sub-section again this year, slightly improved over last year.

Example of review of an Outstanding Section D

"Findings and Conclusions are extremely well articulated, with evidence being well marshalled and a large number of sources being used. Careful language is used to explain the plausibility of how the evidence has been interpreted in most cases. The structuring of the report is also novel, combining background information with findings. This works particularly well in regard to humanitarian response, where the findings relate to particular clusters. The conclusions section is also structured clearly according to the evaluation criteria, and offers an additional level of insight."

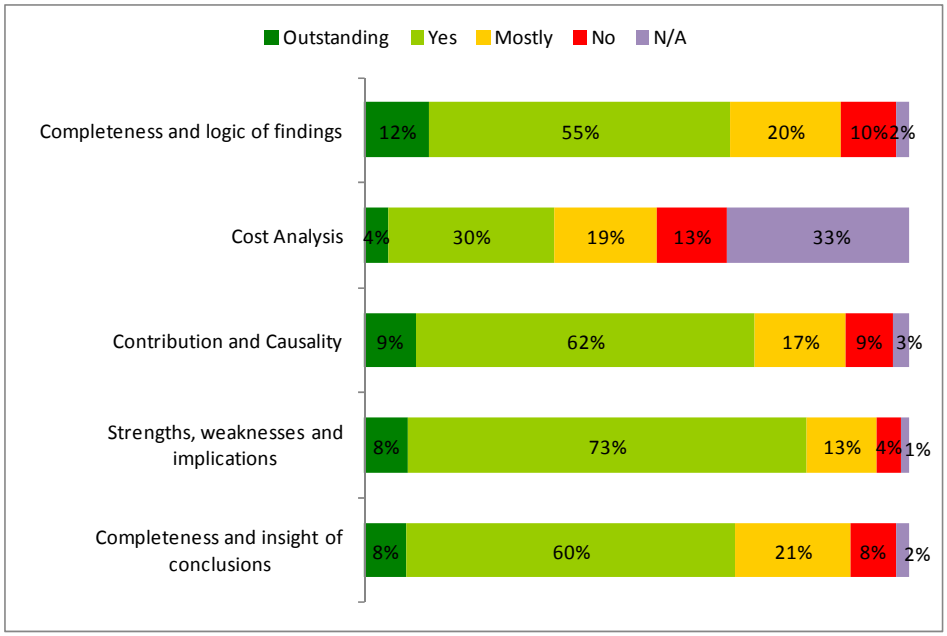
"Real-Time Evaluation of UNICEF's Response to the Typhoon Haiyan in the Philippines" (GEROS-Philippines-2014-004)

Example of review of an Unsatisfactory Section D

"Findings are not presented with supporting evidence from the qualitative or quantitative research… When compared against the evaluation questions…the findings do not…address these questions. The evaluators do not acknowledge the obvious methodological limitations and proceed to evaluate the object for its impact without evidence to substantiate these claims."

"An Evaluation of the Impact of UNICEF Radio Listener Groups Project in Sierra Leone" (GEROS-Sierra-Leone-2014-003)

Exhibit 5.18 Sub-Section Ratings: Findings and Conclusions



In 2014, more than one-third of the reports considered that a **cost analysis** was not applicable in the evaluation, which is higher than in 2013 (when this was considered not to apply in 11% of the reports). This may be due to a variation in interpretation of the template standards, where there is some reviewer discretion involved in terms of whether a cost analysis would be feasible, where it is not specifically requested in a TORs and yet is presented as a standard UNICEF expectation. In any case, cost analysis received the lowest proportion of good quality ratings (34%), which was a decline from 2013, indicating that it is typically not done well and is not improving. Partial points were given for some limited attempt at examining programme costs, cost efficiency, or cost implications for replication or scaling up.

Two other sub-sections improved in 2014. The **completeness and logic of findings** standard refers to findings that are clearly presented, address all criteria and questions, demonstrate a progression to results, and discuss gaps, limitations, and unexpected findings. Good quality ratings improved in 2014 to reach a total of 67%. Reviewers noted gaps in reports that did not cover all the stated evaluation criteria or questions, and did not present an explanation (the most frequent omission was the OECD/DAC criterion of efficiency). Highly rated reports typically structured their findings section in a systematic manner (for example, by evaluation criteria or question) and presented their findings clearly so that they were quickly apparent to the reader (for example, some reports bolded or boxed findings within the text, stated them at the beginning of the section, or separated the findings text from the discussion of evidence).

Improvement was also seen in the **completeness and logic of conclusions** standard, which improved from 2013 to 68% good quality ratings in 2014. This standard asks that a report goes beyond simply restating findings and offers additional insight that adds value, takes account of diverse stakeholders' views, and is appropriately pitched to the end users. While conclusions often do a good job of summarizing findings, reviewer comments indicate that they sometimes struggle to add additional insight and value beyond this summary. Nonetheless, the majority of reports met this standard.

One reoccurring issue in reviewer comments concerns both findings and conclusions of reports. This is the tendency of some reports to create a confusing mix of two or more of findings, conclusions, recommendations, or lessons learned in individual sections of the report, so that it becomes difficult for the reader to locate the pertinent information they are looking for, or to quickly understand the key messages of the report. Reviewers complemented reports that clearly differentiated these types of statements, while linking them to relevant components (for example, recommendations that are clearly linked to findings and conclusions).

Contribution and causality was largely unchanged from last year, with 71% of good quality ratings. This suggests that most reports do a good job of assigning contribution for results to stakeholders, and identifying causal reasons for accomplishments and failures.

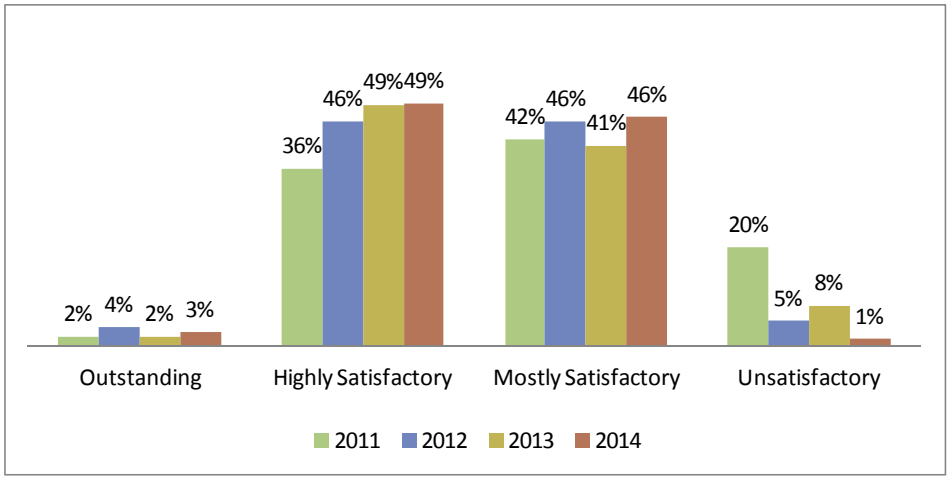
The most highly rated sub-section was **strengths, weaknesses, and implications**, with 81% good quality ratings. This was an improvement from last year. Reviewers commended reports for discussing future implications of constraints, and offering a balanced presentation of the strengths and weaknesses of the evaluated object.

5.4.6 Section E: Recommendations and Lessons Learned

Finding 12: Largely unchanged since 2014, the number of good quality reports for Section E remained the lowest of all sections. Clearer identification of target stakeholder groups and lessons learned could help improve the ratings for this section.

Good quality ratings for Section E remain the lowest among all the GEROS template sections, as in previous years. There has been a general upward trend over the last several years (Exhibit 5.19). The most notable change from 2013 for this section is a decrease in unsatisfactory reports, from 8% to 1.

Exhibit 5.19 Section E Ratings and Comparison with (2011-2014)



Based on sub-section ratings (Exhibit 5.20), the **relevance and clarity of recommendations** continued to be the strongest component by far of Section E at 75% good quality ratings, which is nearly identical to last year. Reviewers commended reports for producing recommendations relevant to the purpose of the evaluation, and grounding them in the evidence presented. However, there was a clear weakness in prioritising recommendations, an individual indicator that received only 50% good quality ratings. Reviewers often criticised reports for producing too many recommendations, or those that were vague or overly-specific.

Example of review of an Outstanding Section E

"The recommendations and lessons learned are very relevant and actionable to the purpose and objectives of the evaluation. They are supported by the findings and were developed with the involvement of relevant stakeholders. They identify the target groups for action and reflect a very good understanding of the country context and UNICEF's role."

"Let Us Learn (LUL)" Formative Evaluation UNICEF Afghanistan Country Office (GEROS-Afghanistan-2014-008)

Example of review of a Mostly Satisfactory Section E

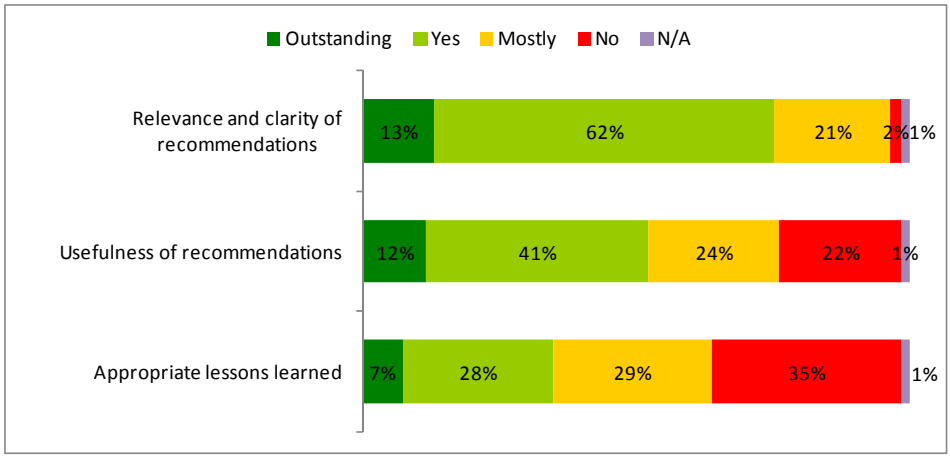
"Les recommandations sont pertinentes et réalistes mais ne sont pas priorisées ce qui limite leur utilité pour les décideurs. Le rapport ne présente pas de leçons tirées en lien avec les objets de l'évaluation."

"Évaluation des Interventions à Base Communautaire dans les Régions des Savanes et de la Kara" (GEROS-Togo-2014-001)

Though the proportion of good quality ratings related to the **usefulness of recommendations** increased since 2013, good quality ratings in this standard apply to just half of reports, and some notable areas of weakness can still be identified. Looking at the individual indicator ratings for this sub-section, only half of reports clearly identified target groups for action, a frequent weakness noted in reviewer comments. While reviewers generally felt the recommendations were realistic, the majority did not describe the process followed in developing the recommendations. At only 30% good quality ratings, this is the most poorly rated individual criterion in the template. Since there is no discussion in most reports of how recommendations were developed, it is uncertain whether the process is lacking participation by important stakeholders, or whether reports are simply lacking a description of the processes they are following. Either way, this does not appear to be a requirement that is well understood by evaluators.

The identification of **appropriate lessons learned** remained, as last year, the most important weakness in Section E, with only 35% of reports rated good quality in this respect. Based on ratings for the individual indicators as well as reviewer comments, about one-third of reports did not include lessons learned at all. Reviewer comments indicate that frequently, there was no attempt to generalise them to indicate their wider relevance beyond the evaluated object. This suggests that many evaluators and/or evaluation managers do not understand UNICEF's definition of lessons learned. Sometimes they are not specifically mentioned in the TORs, and ensuring that they are may assist in improving this practice.

Exhibit 5.20 Sub-Section Ratings: Recommendations and Lessons Learned

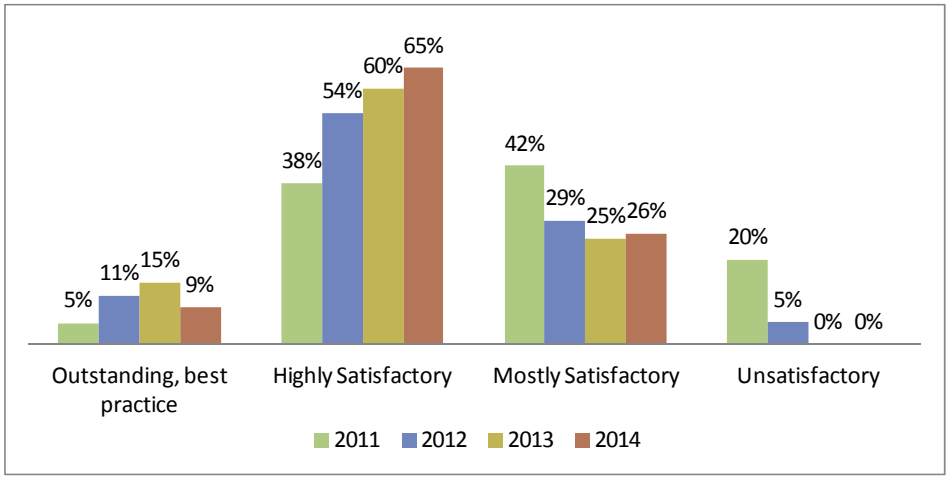


5.4.7 Section F: Report Structure, Logic and Clarity

Finding 13: Good quality ratings for Section F did not significantly change compared to 2013, but this Section continues to have amongst the highest overall good quality ratings. The majority of evaluations were logically structured, but issues were noted regarding the length of executive summaries or their ability to stand alone.

There have been minimal changes in quality of report structure, logic and clarity (Section F) since last year. The proportion of reports considered to be good quality (74%) remained virtually the same as last year. The proportion of reports with lower ratings was virtually unchanged, with none considered unsatisfactory.

Exhibit 5.21 Section F Ratings and Comparison (2011-2014)



In 2014, good quality ratings for **style and presentation** increased slightly to 76%. Reviewers were generally complementary about the inclusion of basic elements in the opening pages of the report, and structuring the report in a logical manner. Often, this reflected a structure based on criteria or evaluation questions. As noted in Section D, however, there were sometimes problems within sections of reports that mixed findings, conclusions, and recommendations in a confusing manner. Annexes were usually found to contain appropriate elements that added credibility to the report, although elements such as data collection tools were sometimes missing.

Example of review of an Outstanding Section F

"This is a very strong report in terms of style and accessibility. It strikes an admirable balance between readability and detail. All requirements of the UNICEF/UNEG standards are met in terms of the elements to be included."

Real-Time Evaluation of UNICEF's Response to the Typhoon Haiyan in the Philippines (GEROS-Philippines-2014-004)

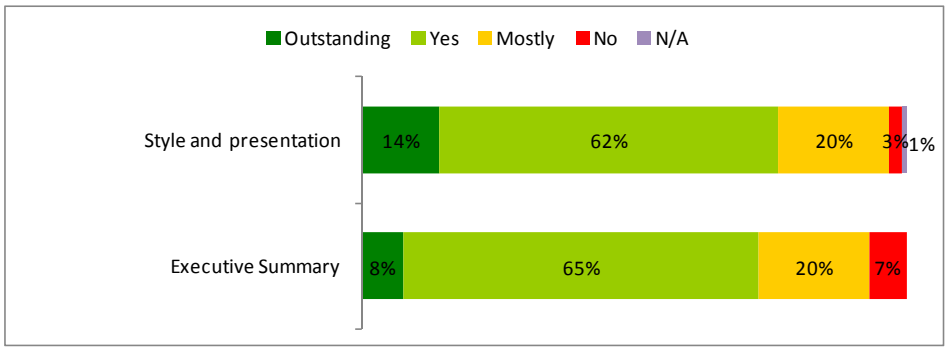
An **Executive Summary** was included in nearly all reports, and received 73% good quality ratings. However, the individual standards ratings show that reviewers found some deficiencies in the ability of these summaries to stand alone (i.e. to not require reference to the rest of the report) and to inform decision making. Both of these individual indicators received 61% good quality ratings, and comments frequently reflect a concern with the long length of executive summaries, compared to UNICEF's suggestion in the GEROS template of 2-3 pages. In most cases, executive summaries included all the expected elements.

Example of review of a Mostly Satisfactory Section F

"The general outline of the report is logically structured…But the… large amount of repetition of information decreases the report's readability and leaves findings unclear... The executive summary provides a reasonable summary and is an appropriate length, but still does not contain an overall higher-level analysis."

"Impact Evaluation of Water, Sanitation and Hygiene (WASH) within the UNICEF Country Programme of Cooperation, Government of Nigeria and UNICEF, 2009-2013" (GEROS-Nigeria-2014-012)

Exhibit 5.22 Sub-Section Ratings: Report Structure, Logic and Clarity



5.4.8 Coherence

Finding 14: Three-quarters of reports reviewed in 2014 did a good job of providing a coherent overall narrative.

In order for a report to be coherent, according to GEROS standards, it must be consistent and logical, flowing clearly from one section to the next. In 2014, 81% of reports were deemed to be coherent. This is similar to last year's assessment, which was an improvement from 58% in 2012. Incoherent reports were few and were at the same level as 2013 (3%). In cases where reports lacked coherence this was often due to insufficiently linking elements of the report (findings, conclusions, and recommendations), leaving out key elements of the evaluation (conclusions, lessons, recommendations), or having a large number of typos and other editing issues that affected readability of the report. .

5.4.9 Addressing the gaps in Evaluation Reports

Based on the analysis of the trends by section, there are a number of areas that require further attention in order to improve the quality of UNICEF evaluation reports.

- Section A: Theories of change are often missing or not properly developed in evaluation reports. This is understandable since this is a fairly new evaluation approach and not all evaluators or UNICEF officers are entirely comfortable with this concept. TORs should clearly specify when a theory of change needs to be developed or refined. In cases where this is not applicable or impossible, evaluation reports should state the reasons why.
- Section B: One of the GEROS standards requires a justification of the evaluation criteria that are used, or not used, in the evaluation. In the reports reviewed, there is often no rationale provided for the use of specific evaluation criteria. Evaluators may not be aware of the requirement to justify evaluation criteria, especially when those criteria are first identified in the TORs and/or when the evaluator is using standard OECD/DAC evaluation criteria.
- Section C: Evaluations reviewed often lacked a counterfactual to address issues of contribution or attribution. However, in many cases, a counterfactual may be impractical or not relevant. This is an area for more specific guidance in TORs and dialogue between UNICEF and the evaluators about when it may or may not be appropriate. Another element which is often poorly executed in the reports is the integration of gender equality. UNICEF has been reporting on the Evaluation Indicator of the UN System-Wide Action Plan on gender equality and women's empowerment. More dissemination of the UN-SWAP criteria for gender responsive evaluation is required since country offices are not all aware of it. Gender responsive evaluation begins from the development of the TORs and UNICEF officers have a great deal of responsibility in ensuring that this is properly done.
- Section D: UNICEF expects that a cost analysis will be carried out as part of the evaluation. and yet it is rarely done and seldom included in TORs. Costs analyses are not always possible, but when they are, clear expectations should be included in the TORs.
- Section E: The identification of lessons learned is a weak area, mostly because lessons identified in evaluation reports are not generalizable to indicate their wider relevance beyond the evaluated object. This seems to indicate that many evaluators and/or evaluation managers do not understand UNICEF's definition of lessons learned. Sometimes they are not specifically mentioned in the TORs, and ensuring that they are may assist in improving this practice.

- Section F: To be useful for decision makers, executive summaries need to stand alone (i.e. to not require reference to the rest of the report), and yet, they also have to be concise. Executive summaries are often completed only when the final version of the report is submitted to UNICEF, which does not leave room for comments or revisions. Evaluation managers should carefully review these summaries to ensure that they can be useful.

5.5 Examples of Good Practices

The GEROS assessments were examined with a view to draw out examples of good practices in different evaluation reports. In the table below, some of these practices are identified:

Section A: Object of the Evaluation	Including both a narrative and a schematic description of the project Theory of Change (TOC) clearly linking project outputs to outcomes is a good practice that should be encouraged. Theories of change are not only useful tools in identifying what a program can influence, but they also provide insights on whether programs can reach their goals with the time and resources they had available. Oftentimes, when a TOC is well explained at the beginning of the report, it is easier to follow the logic of the progression to results in the report. There is a clear distinction between results at different levels and evidence is used throughout to support judgements.
Section B: Evaluation purpose, objectives and scope	Having the evaluators take the description and justification of the evaluation criteria a step further by explaining how they intend to interpret and evaluate each criterion was identified as a good practice.
Section C: Evaluation Methodology and Gender, Human Rights and Equity	Reports that described methodological choices in a clear and transparent way and provided detailed descriptive elements were considered to be good practice. Effort made by the evaluation team to overcome the limitation of not being able to construct a true counterfactual through creating an ex post defacto baseline was another example of good practice. Including considerations for cross-cutting issues, in particular gender equality & human rights and equity, consistently throughout the evaluation is also a good practice. Using human rights vocabulary throughout the report and clearly identifying stakeholders as rights holders and/or duty bearers is another way to ensure proper use of HRBA, and so is including references to the international human rights framework as well as national rights benchmarks.
Section D: Findings and Conclusions	Ideally, the organization of the findings should be based on the established evaluation criteria. Providing definitions of each criterion so that the reader clearly understands how that criterion was used in the report is a good practice. Good practice was also visible in reports that clearly distinguished between findings supported by evidence and conclusions derived from evaluator assessment of the findings.

Section E: Recommendations and Lessons Learned	Including a breakdown of recommendations into strategic and tactical, an indicative timeframe for each, as well as a summative cost-benefit analysis for each demonstrates a good understanding of UNICEF capabilities and processes. Another good practice noted related to recommendation was the inclusion of a table for each recommendation which summarizes the recommendation, describes the action required as well as the specific 'task holder'.
Section F: Report is well structured, logic and clear	Data visualisation is important and extra effort to display the data in a manner that is convenient for the reader is another good practice. The adequate use of summary tables, charts and graphs can greatly improve the readability of a report.

6 Conclusions

The following conclusions are derived from the analysis provided in the present report as well as discussions held with the review team over the course of the 2014 GEROS exercise.

The quality of UNICEF's evaluation reports continued to increase overall through 2014, but only moderately.

Though the rates of increase in quality are declining in recent years, the upward trend in report quality maintained itself for another year. Good quality ratings rose from 69% to 74%. There are a few factors that may be contributing to this increase this year. Firstly, at the regional level, reports from some regions increased or maintained their level of good quality ratings (HQ, CEE-CIS, and ROSA) while others declined in quality (ESARO, WCARO). Those that declined in quality also submitted many fewer reports than in 2013. This may have had an impact on overall GEROS quality ratings, and also raises the question of whether there are links between the quality of evaluations and the extent of participation in the GEROS process. Secondly, a far greater proportion of reports included TORs this year, from 57% to 88%. This may also be a contributing factor in improved quality, as in previous years, reports with TORs tended to be rated higher than reports without TORs.

The contribution of reports to the GEROS process by some UNICEF's regional and country offices appears to be dropping.

In 2014, the total number of reports submitted to GEROS declined from 96 reports in 2013 to 69. This decline reflects many fewer reports submitted by LACRO, ESARO, and MENARO, while other regions (ROSA) and HQ submitted more. This may reflect diverging interest from regions in participating in the GEROS process. These changes have also resulted in less linguistic diversity in reports being submitted.

Certain SPOA areas are covered more than others (i.e. education, which constitutes 32% of evaluations), and a few areas may be under-represented in the overall mix. Programme and project evaluations are by far the most common, and country programme evaluations remain few in number (3), despite the importance of the country as a unit of analysis. UNICEF may want to know more about the issue of evaluation coverage and eventually complement the information obtained through GEROS in order to determine if it is achieving an acceptable level of coverage.

Reports continue to demonstrate similar shortcomings found in 2013.

The data collected during this review cycle revealed that, while good quality ratings for most of the sub-sections in the GEROS template improved, some standards remained particularly weak, especially those relating to ethics, theory of change, justifying evaluation criteria, cost analysis, human rights and gender, targeting recommendations (and explaining the process used to develop recommendations), and lessons learned. In some cases, these can be addressed through filling gaps in the description of certain elements which may have been present in the background of the evaluation (such as ethical safeguards and a theory of change), or making explicit that which has been left implicit in the narrative. In other cases, a significant improvement will require different approaches and analysis, as in conducting cost analysis. It appears that many evaluators do not understand UNICEF's definition or expectations around some components, such as lessons learned. Though not all these components are paramount to methodological robustness, they remain important to UNICEF and their omission/lack of detail had negative repercussions on overall performance in the GEROS process. On the other hand, the far greater inclusion of the TORs in reports this year was a positive development and also aided reviewers in more accurately assessing GEROS ratings.

While the majority of reports are increasing in quality, considerable improvement is now most likely to be achieved through careful analysis of evaluation management structures and oversight systems. Clear, complete and detailed TORs favour the production of better workplans, which in turn enhance the quality of drafts and evaluation reports. The management and oversight of each evaluation stage should incorporate, communicate and apply relevant UNICEF standards. According to the 2014 *Annual Report on the Evaluation Function and Major Evaluations*, regions are already moving in this direction.

After six years of applying the GEROS template with minor changes, it may be time to more significantly adjust the content and structure of the template to address ongoing issues.

While small adjustments have been made over time to improve the GEROS template, some areas of the template have proven to be problematic, in terms of maintaining consistency in interpretation and ratings, efficiency, or allowing flexibility for different types of evaluations. Improvements could be made to the rating system, the structure and content of the framework, and the GEROS process. Details of problematic areas and suggested changes, based on Universalia's three years of experience with this process, have been addressed in Appendix III.

7 Recommendations

As in previous years, the recommendations provided herewith aim to encourage the improvement of UNICEF's broader evaluation practices as well as the quality of its evaluations, through the effective application of GEROS as a quality assurance system.

The following recommendations were formulated in view of the findings and conclusions of this report, based on discussions held with the GEROS review team at Universalia. At the request of UNICEF, Universalia has also made specific comments and suggestions for changes to the GEROS template (see Appendix III). These recommendations incorporate feedback from the UNICEF Evaluation Office and Regional Offices on the draft report.

Recommendation 1: UNICEF should examine whether the increase in quality of evaluation reports, as assessed through GEROS, has resulted in senior managers having greater confidence in evaluation reports.

The GEROS system is an important element in harmonising evaluation standards across UNICEF evaluations, at the global level. Indeed, the GEROS exercises conducted thus far have noted a steady increase in the quality of reports submitted, which should also mean that they are better meeting the needs of decision-makers. The GEROS process – and evaluations more generally – remains only one piece of a larger, organisation-wide feedback system.

Given that the quality of decentralised evaluation appears to be on the right track, it is now time to think about and assess the overall feedback system that is available to senior management, and how evaluation fits into that system. Going forward, efforts should be made to analyse how the GEROS system contributes to the increased use and value of UNICEF's evaluation function, especially at the regional and country levels. More specifically, UNICEF should see if, as a result of the increased reporting quality noted, senior managers now have greater confidence in the reports produced and whether the information obtained is useable. In addition, the EO may want to further understand the reasons for the decline in evaluations submitted to GEROS from several regions in 2014. It may be that regions are tending to use other types of feedback (e.g., from research) more often than evaluations.

This recommendation is relevant to the Evaluation Office.

Recommendation 2: Within its decentralised evaluation strategy, UNICEF should continue to build its own regional/country office evaluation capacities and national capacities to conduct relevant types of evaluations.

UNICEF regions play a role in strengthening evaluation within UNICEF. While good foundations have been established, there is still a need to ensure that targeted, focused support is provided to build UNICEF regional and country office capacity and national evaluation capacities to address recurrent shortcomings. This capacity support should take into account a number of issues, including:

- The majority of evaluation reports submitted in 2014 focused at the programme and project levels. However, is there a need for more evaluation at a strategic level, including evaluation of upstream work, cross-cutting themes, etc?
- If the country is a critical unit of analysis in development (and important to inform future Country Programme documents), UNICEF should continue to encourage country-level evaluations (including country programme evaluations or joint-evaluations).

- UNICEF has encouraged country-led evaluations, where it acts as a partner rather than a leader in the evaluation process, but there are still relatively few of these being carried out. UNICEF should continue to promote and support the development of capacity for country-led evaluation, which is likely to be an important element in implementing the Post-2015 Agenda.

This recommendation is relevant to the Evaluation Office and Regional Offices.

Recommendation 3: Special efforts should be made to strengthen certain aspects of evaluation reports that have been consistently weak in the past few years.

Over the past few GEROS cycles, performance on some evaluation components has been consistently weak. Indeed, the description of the theory of change, ethical considerations, and development of targeted recommendations; the justification for evaluation criteria; the integration of gender and human rights; as well as the systematic inclusion of good quality lessons learned have been challenging. Cost analysis is another component that requires greater attention or needs to be reframed in a revised GEROS review framework, if it continues to be important to UNICEF. In order to address these weaknesses, UNICEF should focus further efforts and attention on improved training and guidance around these standards in particular.

This recommendation is relevant to the Evaluation Office.

Recommendation 4: UNICEF should continue to update and systematically communicate its requirements for evaluation reports across its entire evaluation oversight/management system. These updates should take into account evolving standards for evaluation in the UN System.

In an effort to continuously improve its evaluation oversight system, UNICEF should ensure that its priorities, standards and criteria are clear and up-to-date across all evaluation management levels, including whether it best reflects UNICEF's and UN requirements, the current state of practice, and a streamlined review process. This includes reviewing and clarifying the GEROS template (which UNICEF does periodically) as well as the systematic integration of evaluation priorities and standards in all TORs, which should translate into better quality inception reports and thus, improved evaluation reports. This would also include the UN-SWAP standards which is not widely known in countries. For their part, Regional Offices can follow up with Country Offices to emphasise the importance of systematically including TORs within draft and final evaluation reports. Specific suggestions on amendments to the GEROS template is provided in Appendix III.

This recommendation is relevant to the Evaluation Office and Regional Offices.

Recommendation 5: As part of the periodic review of GEROS, UNICEF should consider revising the rating scale and several elements of the GEROS template in order to ensure greater precision in the messages that are provided about evaluation quality and the characteristics of evaluation reports, and to create more efficiency in applying the template.

This recommendation emerges from the experience of conducting this meta-evaluation over the past three years. For the 2012 cycle, after three years of implementing GEROS, UNICEF changed the rating scale. The rating of "Almost Satisfactory" was changed to "Mostly Satisfactory", while the traffic light color (yellow) associated with the rating remained the same. The use of the term "Mostly Satisfactory" for a yellow rating may be misleading, as a yellow traffic light suggests something in which there is only partial compliance, and where there is still room to improve in order to become satisfactory. In addition, the change in the rating also created a larger gap between "Unsatisfactory" and "Mostly Satisfactory" ratings, which creates challenges in making the judgment about how to rate particular questions or standards in the template. In Appendix III, we provide further comments on the scale and identified other areas where greater clarity or efficiency can be introduced in applying the evaluation report classification and quality standards, as described in the GEROS template.

8 Lessons Learned

UNICEF defines lessons learned as contributions to general knowledge that help expand common understanding. The following lessons aim to build on the lessons provided in the 2013 meta-analysis, based on the experience of the 2014 review process.

Clear and systematic communication of evaluation standards and priorities favours the effective alignment of evaluations with UNICEF standards, from the outset (i.e. TORs stage).

The terms of reference constitute the key starting document of any evaluation. In order to ensure that evaluations respond to UNICEF's expectations and priorities, evaluation TORs should clearly identify the standards by which reports will be judged in the GEROS process. In this way, high quality reports will not be penalised for omitting certain components due to a lack of awareness of UNICEF expectations both on the part of the evaluator and the evaluation manager.

While common standards help improve evaluation quality, quality assurance systems such as GEROS should provide sufficient flexibility to account for different types of evaluations.

Common standards help ensure that evaluation reports are judged according to the same standards and strive to attain specific goals and objectives. However, a balance should be found between common standards and flexibility, to take into account the different types of evaluations produced each year within UNICEF. Currently, the GEROS process illustrates this tension between set objectives and flexibility, as the template used does not always easily adapt itself to specific types of evaluation reports (e.g. impact evaluations, and separately published case study evaluations which are part of a broader evaluation).

In a decentralized system, compliance with quality assurance systems such as GEROS is affected by incentives, available resources, and the perception of relevance.

UNICEF's decentralized structure of regional and country offices leads to challenges in maintaining consistency in how evaluations are conducted, a challenge which GEROS was created to assist with. For full participation in a process such as GEROS, the relevance and utility of the process must be clear to the regions. Part of maintaining relevance is ensuring that the system is continuously updated to reflect changing expectations and priorities.

Quality assurance systems such as GEROS need to strike a balance between consistent application over a period of time (which allows for comparison) and making major adjustments in order to improve utility and reflect changes in the environment.

Updating and changing the GEROS template every year is not necessary and would be time consuming. It is however necessary to have a formal and consultative review every two to three years in order to keep the process relevant and to update the GEROS template as required in light of new developments or requirements.

Appendix I List of Findings

Finding 1: In 2014, the quality of reports submitted continued to increase overall, accompanied by a large increase in the number of reports with TORs included in the appendices or as separate documents.

Finding 2: The quality of UNICEF evaluation reports varied since last year's review process, with some regions improving and others declining.

Finding 3: As in previous years, most reports focused on initiatives with a national scope and they were overall good quality reports.

Finding 4: Most evaluation reports in the sample were managed by UNICEF and were considered quality reports. With respect to evaluation purpose, programme and project evaluations continue to represent the most important proportion of evaluations reviewed.

Finding 5: A larger proportion of evaluations are summative this year and a majority of reports focus on education as a thematic area.

Finding 6: Almost all reports submitted to GEROS are independent external evaluations, which is a shift since 2012 when internal evaluations (or reviews) were still sometimes classified as evaluations. A majority of the reports submitted were in English.

Finding 7: In 2014, the quality of reports slightly increased in two sections, decreased slightly in another, and changed little in three other sections. The majority of reports continue to be aligned with the UNICEF-adapted UNEG standards for evaluation reports.

Finding 8: In 2014, the description of the evaluated object and its context declined somewhat in quality compared to 2013. Evidence suggests that the description of the theory of change and of stakeholder roles and contributions remain areas for improvement.

Finding 9: The extent to which reports met standards of evaluation purpose, objectives and scope changed little from previous years. Strengths of reports in this area included clear purpose, objectives and scope and a clear list of evaluation criteria. However, the justification for the selection of evaluation criteria still requires greater attention.

Finding 10: Reports have made improvements in terms of the description of the methodology. While methodological robustness is often satisfactory, lacunas in ethical considerations and stakeholder participation may have impacted the overall ratings of this section.

Finding 11: The presentation of findings, conclusions, contribution and causality continued to improve in 2014. Cost analysis remained particularly challenging.

Finding 12: Largely unchanged since 2014, the number of good quality reports for Section E remained the lowest of all sections. Clearer identification of target stakeholder groups and lessons learned could help improve the ratings for this section.

Finding 13: Good quality ratings for Section F did not significantly change compared to 2013, but this Section continues to have amongst the highest overall good quality ratings. The majority of evaluations were logically structured, but issues were noted regarding the length of executive summaries or their ability to stand alone.

Finding 14: Three-quarters of reports reviewed in 2014 did a good job of providing a coherent overall narrative.

Appendix II List of Recommendations

Recommendation 1:	UNICEF should examine whether the increase in quality of evaluation reports, as assessed through GEROS, has resulted in senior managers having greater confidence in evaluation reports.
Recommendation 2:	Within its decentralised evaluation strategy, UNICEF should continue to build its own regional/country office evaluation capacities and national capacities to conduct relevant types of evaluations.
Recommendation 3:	Special efforts should be made to strengthen certain aspects of evaluation reports that have been consistently weak in the past few years.
Recommendation 4:	UNICEF should continue to update and systematically communicate its requirements for evaluation reports across its entire evaluation oversight/management system. These updates should take into account evolving standards for evaluation in the UN System.
Recommendation 5:	As part of the periodic review of GEROS, UNICEF should consider revising the rating scale and several elements of the GEROS template in order to ensure greater precision in the messages that are provided about evaluation quality and the characteristics of evaluation reports, and to create more efficiency in applying the template.

Appendix III Detailed Comments on GEROS Framework

Based on Universalia's multi-year experience as the external consulting firm managing the GEROS process, we offer these comments to contribute to future improvements of GEROS in terms of improving efficiency and rating consistency. These considerations could be used to update the template and the accompanying process for the next GEROS cycle.

Iterative structure of the template

From a user perspective, the structure and design of the template works well in terms of organizing a large number of criteria into clear divisions and visually demarcating which ratings have been applied. However, we find that the cascading structure of multiple columns (plus a row) for comments on criteria is unnecessarily repetitive and might be made more efficient without loss of important data. In the current template:

- Of the comment columns in the spreadsheet, the "Remarks" column provides the details and justification for each rating.
- The "constructive feedback" column provides recommendations on how to improve the report.
- The Executive Feedback row at the end of each section provides, at a glance, the strengths and weaknesses of the report for managerial purposes.
- Column J is intended to summarize comments on all indicators in the section, particularly focusing on the cornerstone questions.

We find that Column J is an unnecessary level of summary that does not contribute in a significant way to further analysis. It is additionally confusing to reviewers in that the template annotations in Column J emphasize the summary of only certain criteria (the cornerstone questions), leading to inconsistencies in how reviewers complete this column. **We would suggest the removal of this column, and allow the Executive Feedback row to act as the summary for the section**.

Rating scale

There are several interconnected issues we would like to identify with the 4-point rating scale (5 if Not Applicable is included) used in the GEROS template. We recommend that UNICEF consider revising the rating scales and adding a few clarifications.

- **The current rating scale does not help to differentiate levels of quality**: While the Outstanding, Highly Satisfactory, and Mostly Satisfactory levels are separated from one another in what feels like equivalent increments, there is a large gap between the ratings of Mostly Satisfactory and Unsatisfactory. This gap is partly due to the phrase" Mostly Satisfactory" itself (which suggests that the report is doing OK), yet it really signals only partial compliance with a particular standard. This has created difficulty for reviewers in the overall rating of a report, where a report is not completely Unsatisfactory, but is also not Mostly Satisfactory. In the end, judgment must be made about what rating is the best fit for that particular report, yet the current rating scale does not facilitate that.

- **Rating colors and labels**: While the use of dark green, light green, amber, and red intuitively reflect a continuum of positive and negative, the use of amber for Mostly Satisfactory is not an intuitive match. With the label suggesting a positive, and the color suggesting problems, this mismatch contributes to inconsistency between reviewers in how the rating scale is interpreted and applied. It also misleads the "readers" or "recipients" of the review, who might celebrate a "Mostly Satisfactory," when in fact it reflects that there is room for significant improvement.
- **Use of the Unsatisfactory rating:** The absence of a rating between Unsatisfactory and Mostly Satisfactory also causes challenges for rating individual questions. With no rating in between these, nor a rating to differentiate 'absent' from something that is present but not well done, reviewers have typically reserved Unsatisfactory (red) for questions where the standard is completely absent or very poorly addressed, and Mostly Satisfactory if the report partially meets the standard. However, in the future it may be necessary to distinguish reports in which the standard is completely absent and to further refine the meaning of partial compliance.
- **Use of Not Applicable**: It has been an on-going challenge to decide when to apply a rating of Not Applicable rather than Unsatisfactory. At times, reviewers have used Not Applicable to indicate the absence of something. For example, if no cost analysis has been conducted, reviewers have sometimes rated criteria 35 ("Is a cost analysis presented that is well grounded in the findings reported?") as Not Applicable, as it was not mentioned in the TORs, and also may have been seen by reviewers to not be feasible for a particular evaluation. In other cases, its absence was rated Unsatisfactory, based on the interpretation that it is a UNICEF expectation regardless of whether it is in the TORs or not, and its absence should have been justified. We managed this by discussing each review on a case-by-case basis, however, in the future, clear parameters for use of 'Not Applicable' should be established with UNICEF, in order to further enhance consistency in the ratings.

Content of the template

The six sections of the GEROS template each contain between five and 13 questions, grouping together related criteria.

Section C: Methodology: While in most cases the criteria are logically organized, Section C (methodology, gender, human rights, and equity) contains disparate elements and an unwieldy number of criteria (13), making it difficult to manage and summarize. The largest grouping of criteria within this section, related to human rights, gender, and equity, could be removed and put in its own section to balance the size. **We would suggest having a specific section on cross-cutting issues and/or gender-responsive evaluation in the template that could also include components of the SWAP.** This consolidation of the two templates, as discussed below, would simplify the overall GEROS process. **If this were done, the existing gender-related criteria should be amended to eliminate the significant overlap with the SWAP criteria.**

Section C: Gender, human rights, and equity: This sub-section of Section C poses challenges for reviewers. Criteria 20 and 21 conflate the three themes of gender, human rights, and equity together, such that a report that addresses one theme well might receive a positive rating while neglecting the other themes to varying extents. Further, there is considerable overlap between criteria 20-24, leaving little for a reviewer to differentiate between 20 and 21, or compared to 22-24. It is time-consuming for a reviewer to search for the subtleties between these criteria ratings in a report, while adding little to the overall analysis. **Having a distinct section on cross-cutting issues could allow**

UNICEF to expand on the gender equality, human rights and equity themes as opposed to having them amalgamated into the same criteria.

Comments on other criteria:

Specific criteria pose particular challenges to reviewers. These are summarized in the following table:

Criteria	Issue and comment	Possible action
3. Does [the context] illuminate findings?	This criterion does not add significant value to the analysis. Well-explained context that is related to the object (criteria 2) naturally illuminates findings.	Eliminate criteria 3.
12. Does the evaluation provide a relevant list of evaluation criteria that are explicitly justified as appropriate for the Purpose?	Due to the phrasing in the template, there has been confusion as to whether the standard OECD/DAC criteria still requires some kind of justification or explanation. We have provided internal guidance to reviewers, but this could be made more explicit in the template itself.	Be explicit in the template about whether standard criteria require any explanation or justification.
13. Does the evaluation explain why the evaluation criteria were chosen and/or any standard DAC evaluation criteria rejected?	This criterion overlaps with #12, and it is unclear how it should be rated if all the evaluation criteria used are those of the OECD/DAC and are not justified.	Be explicit in the template about whether standard criteria require any explanation or justification, and/or, amalgamate with #12.
16. Are ethical issues and considerations described?	The annotation for this criterion suggests that the design of the evaluation should contemplate how ethical the initial design of the programme was, as well as the ethics of how is included and excluded in the evaluation. It is very uncommon that any report would address the first point, contributing to unnecessarily low ratings on this criterion. The issue of participation in the evaluation is a significantly different concept and is also addressed in criterion #26.	Target the content of this indicator more specifically or consider moving this indicator to the Stakeholder Participation sub-section.
20-24 Human Rights, Gender and Equity sub-section	See discussion above.	
35. Is a cost analysis presented that is well grounded in the findings reported?	It is not clear whether a cost analysis should be expected from all evaluation reports or only when ToRs request such an analysis.	Clarify whether a cost analysis is an obligatory or optional component of an evaluation report.
49. Are lessons learned correctly identified? 50. Are lessons learned generalised to indicate what wider relevance they may have?	These criteria overlap. If lessons learned are correctly identified, it necessitates that they are generalisable, since that is the definition of a lesson learned.	Clarify the wording to differentiate these criteria or amalgamate them.
55. Is an executive summary included as part of the report?	This is worded as a Yes/No question, but the reviewer is required to select from the 4-point rating scale.	Change this rating to Yes/No.

Background section (Types of Reports): The top portion of the GEROS template asks reviewers to classify reports into a number of categories (geographic scope, level of results, etc). This provides a useful snapshot during the meta-analysis of the types of reports included in the sample and how they performed. However, some of these categories are problematic, in that they are not always made clear in the report, there is insufficient guidance in the template for that category, or that categories overlap. Specific challenges include:

- The Management category: it is not always clear from evaluation reports which Management category applies, particularly in terms of the level of control that was held by stakeholders outside UNICEF.
- Similarly, the Level of Independence is not always made clear in evaluation reports, particularly the level of control of activities by UNICEF staff or Evaluation Office professionals.
- The inclusion of both Real-time and Humanitarian as mutually exclusive categories under Purpose. Virtually all Real-time evaluations are Humanitarian evaluations; which category should take precedence? The solution could be to simply put "Other Humanitarian", that would give the room to classify humanitarian evaluations that may not be referred to as Real-Time Evaluations per se.
- The inclusion of Humanitarian as a category under Purpose, as well as a category under SPOA Correspondence. This results in some Humanitarian evaluations being marked as Humanitarian under Purpose, but possibly as something else (eg. WASH) under SPOA Correspondence. While this may accurately reflect a Humanitarian evaluation that focused on a WASH intervention, the meta-analysis of SPOA Correspondence does not reflect this WASH evaluation example in its analysis of Humanitarian evaluations.

As it is not always entirely obvious from a report which classification should be applied, **we suggest that regions submitting the report specifically provide the classification information required using a simple form that should accompany the report.**

Application of the template to different types of evaluations

We have found that the GEROS template leaves little flexibility for other evaluations that are not considered typical, such as impact evaluations or case studies of a larger evaluation study. UNICEF could adapt its current template to account for possible variation in content, structure and focus within reports. Naturally, caution should be taken to prevent the template from becoming so flexible that common standards are easy to circumvent. We offer below some feedback.

- **Case Studies: The overall report is the only one that we recommend be judged**. Often, the country case studies are not designed to be stand-alone evaluations; they are designed as input to a corporate or regional evaluation. Plus, case studies vary significantly in terms of scope and approach. This is why we would argue that they should not be assessed through the GEROS process. However the case studies must be available for review in order to provide enough depth to the analysis.
- **Impact Evaluations**: Impact evaluations are different in a number of respects and in many cases do not follow the typical methodology used by other programme evaluations. They tend to be rated less favorably on cross-cutting issues, for example, or on the application of OECD/DAC criteria. **Ideally, impact evaluations should use a distinct template. If this is not possible, great care and flexibility needs to be used in reviewing one of these reports.**

- **UNDAF Evaluations**: UNICEF will need to decide if it would like to assess UNDAF evaluations as part of the GEROS process. Flexibility is needed to review the UN reports. Most of the criteria in the GEROS template apply to UNICEF evaluation reports, and UN-wide reports are not expected to comply with all these requirements. **If they remain a part of the GEROS process, we recommend that the review of UNDAF or joint-evaluations be assigned to senior reviewers who, based on their experience, can more readily recognize how and whether to rate certain categories in unusual contexts.**

We do not recommend creating a new template for each different type of report since it would diminish the comparability of all the reports for meta-analysis purposes. However, UNICEF must understand that leaving the template as is may lead to lower rating due to the absence/special treatment of some criteria in different types of evaluation reports.

Recommendations for the GEROS template

Re-visit the rating scale and resolve the discrepancies between Mostly Satisfactory and Unsatisfactory. It is the lower end of the rating scale that poses difficulties for reviewers. Changes to the rating scale to address the issues described should include adding a rating level between Unsatisfactory and Mostly Satisfactory (e.g. Partially Satisfactory). In all cases, we suggest that the service provider draft specific guidance on these issues and discuss it with UNICEF, in order to clarify the circumstances in which to use these ratings. As noted in the recommendation on the GEROS process, this kind of guidance or manual should be costed in a future long term agreement with a service provider.

Develop and share definitions of the rating scale options to ensure that UNICEF and reviewers have the same understanding. Clear definitions of what each rating scale option means and when it should be applied should be developed by UNICEF and provided to reviewers.

Separate the human rights, gender and equity sub-section from Section C and place it in its own section. Consideration could also be made of reducing the overlap between these criteria by reducing the number of criteria or differentiating them further from one another. Currently, statements referring to human rights, gender and equity are lumped together, e.g. question 20: "Did the evaluation design and style consider incorporation of the UN and UNICEF's commitment to a human rights-based approach to programming, to gender equality, and to equity?". Specific rating could be provided on each of these items.

Remove Column J from the template and allow the Executive Feedback row to serve as the summary of the section. Removing Column J reduces effort while not removing any data or analysis, and still allows for a summary in the Executive Feedback row, as well as constructive feedback on particular issues.

Amend problematic standards to allow for more efficiency and consistency. As described in the table in the discussion, there are a number of standards which are either difficult to assess, overlap with other standards, or are unclear.

Simplify the report classification categories or ask country office to complete a short form to identify the correct classification information when submitting the report. The report classification questions and options often overlap and are unclear, e.g. in geographic scope, it is often difficult to differentiate between sub-national and national scopes, in purpose, "humanitarian" and "real-time evaluation" are mutually exclusive, and in the level of results sought (outputs or outcomes), it is often impossible to determine which one is right. The management of the evaluation is another problematic area to cover since terms of reference are not always clear about who will manage the evaluation, and evaluation reports seldom mention who exactly they reported to.

Comments on the SWAP Framework

The SWAP template was considerably simplified for 2014 and has eliminated the problematic criteria that often required information not found in evaluation reports. This has enabled reviewers to respond to the criteria in the vast majority of cases and should help lead to more valid ratings.

Possible amalgamation of SWAP and GEROS templates

Last year, one of the possibilities raised by Universalia was to incorporate the SWAP criteria into the GEROS template, so that each report goes through only one integrated rating process, which would help make the review process more efficient. At its previous length of 13 indicators, the SWAP criteria would have been somewhat unwieldy to integrate into GEROS, but with four criteria, this is entirely feasible. There are a few different approaches that could be used:

1) Place SWAP indicators into the GEROS template as-is, in their own section or in a section dedicated to cross-cutting issues, maintaining their own rating scale, or using that of GEROS.
2) Place SWAP indicators into the GEROS template in their own section, and amend or amalgamate these indicators as well as the existing GEROS gender-related indicators to ensure there is no duplication.

These approaches maintain the ability to assess gender ratings independently, though comparability with the 2014 exercise will depend on the specific changes made to the criteria, if any.

Content of the template

Regardless of whether the SWAP template continues to be an independent document or is merged with the GEROS template, a clarification to the criterion 4 annotation would be beneficial:

Criteria	
Criterion 4: The evaluation Findings, Conclusions and Recommendation reflect a gender analysis	The annotation for this criterion states that "The evaluation report should also provide lessons/challenges/recommendations for conducting gender-responsive evaluation based on the experience of that particular evaluation". This is rarely done, nor is it clear that this would be seen to be within the mandate of a typical evaluation. This might better be described as an exceptional activity rather than an expectation, or removing this wording.

Recommendations for the SWAP template

Consider integrating SWAP criteria into the GEROS template. Integrating these criteria into GEROS could lead to greater efficiency without losing the ability to independently analyse or highlight data related to gender.

Amend wording of annotation for Criterion 4. The wording of the annotation for criterion 4 raises an expectation that may not be within the mandate of an evaluation, except for exceptional cases, leading to unnecessarily low ratings on this criterion.

Appendix IV Terms of Reference

TERMS OF REFERENCE FOR GLOBAL EVALUATION QUALITY OVERSIGHT SYSTEM EVALUATION OFFICE

Background

UNICEF put in place an Evaluation Quality Assurance System to ensure evaluations managed/supported by UNICEF meet quality standards. The system is composed of a) the **Global Evaluation Reports Oversight System (GEROS)**, managed by the Evaluation Office (EO), which rates final evaluation reports commissioned by UNICEF Country Offices (CO), Regional Offices and HQ divisions against the UNEG/UNICEF Evaluation Report Standards, and b) **Regional QA Systems**, managed by Regional Offices (RO), which assess draft ToR and draft Reports.

UNICEF is looking for an institution to ensure the reviewing of and rating the quality of draft ToR, as well as draft and final evaluation reports supported by UNICEF country and regional offices all over the world, as well as HQ divisions.

Expected results

The selected institutions will review draft Tor, as well as draft and final evaluation reports in English, French and Spanish received by the EO and selected ROs (up to a maximum of 200 draft/final reports and 50 draft ToR in one year timeframe), rate them against UNEG/UNICEF standards, write an executive feedback to be sent to the CO concerned, and make analysis of trends, key weaknesses and strengths of UNICEF-managed/supported evaluation reports and ToRs.

Expected deliverables

Within the Global Evaluation Quality Oversight System, the selected Institution will deliver the following outputs:

A. Draft ToR and draft Reports (contract to be managed by Regional Offices)

A1: Draft Evaluation ToR and draft Evaluation reports reviewed, rated and executive feedback sent

UNICEF Country Offices are sending the draft ToR and reports to Regional Office for real-time quality review and practical comments on how to improve them. The institution will carry out such review in maximum 3 working days for the draft ToR and 5 working days for the draft reports. The institution will provide professional and practical feedback according to pre-agreed templates (see hyperlink below for the evaluation reports, and attachment for the ToR).

A2: Regional overview of evaluation draft ToR and draft reports reviewed

The institution will undertake an annual review of feedback provided (see attachment with the example of last year); identify lessons to be learned on evaluation ToR and reports. Will compare these results with those which emerged from the two previous yearly exercises undertaken in the region and will identify lessons to be learned, emerging good practices and actionable recommendations to improve the quality assurance system as well as the quality of Evaluations in the specific region assessed.

A3: Regional Evaluation Help Desk.

The objective is to ensure real-time trouble-shooting and ad hoc technical assistance to UNICEF Country Offices when requested, for instance providing a second review of ToR, specific technical notes, etc. Timing and content of any specific task to be agreed about beforehand with the RO concerned.

B. Final Reports (contract to be managed by EO)

B1: Final Evaluation reports reviewed, rated and executive feedback sent

Download the final reports from the UNICEF Intranet database, and review and rate final Evaluation reports received in English, French and Spanish against UNEG/UNICEF standards using the Feedback Template – both the comprehensive as well as the executive one (available at http://www.unicef.org/evaluation/files/Tool_2012_v2.xlsx), highlighting strengths, weaknesses and recommendations to improve the quality of future evaluations reports**.**

The estimated total number of final evaluation reports to be reviewed will be between a minimum of 50 and a maximum of 150 in one year timeframe, out of which about 80% in English, 15% in French and 5% in Spanish.

Reports must be fully rated and the feedback given within 10 working days of receipt. At times, there may be as many as 20 to be handled within the 10 day period. If reports to be rated within the 10 working days exceed 20, the rating time will be extended.

B2. Global analysis of trends, key weaknesses and strengths of reports reviewed

Every year, produce a Meta-evaluation based on the assessments of all final reports reviewed that year highlighting key trends, key weaknesses and strengths of reports reviewed, including lessons learned and good practices on Evaluation reports, and actionable conclusions and recommendations to improve GEREOS as well as the quality of Evaluation reports. Pls refer to the latest meta-evaluation for your easy reference at http://www.unicef.org/evaldatabase/index_60935.html . A PowerPoint highlighting key issues should also be prepared.

Management of the system

This Long Term Agreement covers a) the Global Evaluation Reports Oversight System (GEROS), managed by the Evaluation Office (EO), which rates final evaluation reports commissioned by UNICEF Country Offices (CO), Regional Offices and HQ divisions against the UNEG/UNICEF Evaluation Report Standards, and b) Regional QA Systems, managed by Regional Offices (RO), which assess draft ToR and draft Reports. However, the Evaluation Offices will raise a contract to cover the GEREOS system only, while Regional Offices (if any) will raise separate contracts to cover the Regional QA Systems.

The contract raised by the EO (covering the GEROS System) will be managed by the Senior Evaluation Specialist, Systemic strengthening, with the support of the Knowledge Management (KM) specialist. The contracts (if any) raised by Regional Offices will be managed by the respective Regional M&E Chiefs.

The selected institution will appoint a project manager who will ensure consistency of rating, quality and timely delivery of expected products, and overall coordination with UNICEF Evaluation Office. The project manager will also provide an update on a monthly basis, which will include a tracking matrix highlighting the status of reviews, ratings and executive feedback. The project manager will be the point of contact with UNICEF for any issues related to this Long Term Agreement.

Please note that, to avoid potential conflict of interest, the following will be applied:

- The company that will win this bid, will not review any ToR, draft and/or final evaluation reports of evaluation conducted by the same company
- The reviewer who rates the final evaluation reports will be different from the reviewer who rates the draft ToR and/or draft report

Qualifications

- Excellent and proved knowledge of evaluation methodologies and approaches
- Proven experience with Quality review of evaluation reports, preferably with UN agencies
- Proven practical professional experience in designing and conducting major evaluations
- Excellent analytical and writing skills in English required. Adequacy in French and Spanish required, with excellence in French and Spanish a strong advantage
- Familiarity with UNEG/UNICEF evaluation standards is an asset
- Sectorial knowledge in Child survival and development and at least other two UNICEF area of intervention (education; HIV/AIDS; Child protection; Social protection) in English language
- Knowledge and expertise of other or similar quality assurance systems will also be an asset
- Proven capacities in managing databases

Duration of contract

- The Long Term Agreement will start 1 October 2012 and will expire 30 August 2015.

Appendix V List of Reports Assessed

No.	Country	Region	Year	Sequence #	Title	Rating
1	Afghanistan	South Asia Regional Office	2015	2014/008	Let Us Learn (LUL) Formative Evaluation UNICEF Afghanistan Country Office	Outstanding, best practice
2	CEE/CIS and Baltic States	Central & Eastern Europe, Commonwealth of Independent States RO	2014	2014/008	RKLA3 multi-country evaluation: increasing access and equity in early childhood education – final evaluation report	Outstanding, best practice
3	Rep of Uzbekistan	Central & Eastern Europe, Commonwealth of Independent States RO	2014	2014/003	Evaluation of Country Programme of Co-operation between Government of Uzbekistan and UNICEF 2010-2014	Outstanding, best practice
4	Republic of Montenegro	Central & Eastern Europe, Commonwealth of Independent States RO	2014	2014/004	Final Evaluation of the "Child Care System Reform"	Outstanding, best practice
5	Afghanistan	Corporate (HQ)	2014	2014/001	Afghanistan Country Case Study: Unicef's Upstream Work in Basic Education and Gender Equality 2003-2012	Highly satisfactory
6	Afghanistan	South Asia Regional Office	2014	2014/007	In-depth evaluation of female literacy programme	Highly satisfactory
7	Algeria	Middle East and North Africa Regional Office	2014	2014/003	Evaluation du programme élargi de vaccination dans les camps de réfugiés sahraouis de Tindouf	Highly satisfactory
8	Bangladesh	South Asia Regional Office	2014	2014/001	Final Evaluation of Basic Education for Hard to Reach Urban Working Children (BEHTRUWC) Project (BEHTRUWC) 2nd Phase 2004-2014	Highly satisfactory
9	Bangladesh	South Asia Regional Office	2014	2014/014	Impact evaluation of the WASH SHEWA-B programme in Bangladesh	Highly satisfactory

No.	Country	Region	Year	Sequence #	Title	Rating
10	Bangladesh	South Asia Regional Office	2015	2014/002	Let Us Learn Formative Evaluation (case study)	Highly satisfactory
11	Barbados	Latin America and Caribbean Regional Office	2014	2014/002	Evaluation of the Pilot of the Early Childhood Health Outreach Program in St Vincent and the Grenadines	Highly satisfactory
12	Benin	West and Central Africa Regional Office	2014	2014/001	Évaluation des espaces enfances au Bénin	Highly satisfactory
13	Bosnia and Herzegovina	Central & Eastern Europe, Commonwealth of Independent States RO	2014	2014/004	Increasing Early Opportunities for Children in Bosnia and Herzegovina	Highly satisfactory
14	Brazil	Corporate (HQ)	2014	2014/001	Brazil Country Case Study: UNICEF's Upstream Work in Basic Education and Gender Equality 2003-2012	Highly satisfactory
15	Cambodia	Corporate (HQ)	2014	2014/001	Country Case Study: Cambodia - UNICEF's Upstream Work in Basic Education and Gender Equality 2003-2012	Highly satisfactory
16	CEE/CIS and Baltic States	Central & Eastern Europe, Commonwealth of Independent States RO	2014	2014/020	Multi-Country Evaluation of Regional Knowledge and Leadership Areas: Including All Children in Quality Learning in CEE/CIS	Highly satisfactory
17	Colombia	Latin America and Caribbean Regional Office	2014	2014-001	Joint Evaluation of the Programs of Canadian Cooperation (2009-2013) and Sweden (2011-2013) Implemented in the Country Office for Colombia of the United Nations Fund for Children	Highly satisfactory
18	Comoros	Eastern and Southern Africa Regional Office	2014	2014/002	Rapport de l'évaluation finale du Cadre des Nations Unies pour l'aide au développement (UNDAF) et de son rôle d'appui à la stratégie de croissance et de réduction de la pauvreté (SCRP) (2008-2014)	Highly satisfactory
19	Denmark	Corporate (HQ)	2014	2014/001	Evaluation of UNICEF Supply Division Emergency Supply Response	Highly satisfactory

No.	Country	Region	Year	Sequence #	Title	Rating
20	Ethiopia	Eastern and Southern Africa Regional Office	2014	2014/050	An Evaluation of the Child-to-Child School Readiness Programme in Ethiopia	Highly satisfactory
21	Guyana	Latin America and Caribbean Regional Office	2014	2014/001	Health and Family Life Education	Highly satisfactory
22	India	South Asia Regional Office	2014	2014/001	Evaluation of Empowering Young Girls and Women in Maharashtra, India	Highly satisfactory
23	Indonesia	East Asia and the Pacific Regional Office	2014	2014/002	Education Sector Response to HIV & AIDS	Highly satisfactory
24	Indonesia	East Asia and the Pacific Regional Office	2014	2014/004	Equity- Focused Formative Evaluation of UNICEF's Engagement in the Decentralization Process in Indonesia	Highly satisfactory
25	Kazakhstan	Central & Eastern Europe, Commonwealth of Independent States RO	2014	2014/003	International consultancy on evaluation of the supported by the Government of Norway programme on developing a sustained and operational ombudsman's child protection mechanism that prevents and responds to child abuse, exploitation and family separation in line with international standards	Highly satisfactory
26	Lesotho	Eastern and Southern Africa Regional Office	2014	2014/001	Child Grand Impact Evaluation	Highly satisfactory
27	Macedonia	Central & Eastern Europe, Commonwealth of Independent States RO	2015	2014/007	Evaluation of the Early Childhood Development Programme in the former Yugoslav Republic of Macedonia	Highly satisfactory
28	Malawi	Eastern and Southern Africa Regional Office	2014	2014/012	Evaluation of PHC Essential Medicines Project	Highly satisfactory
29	Maldives	South Asia Regional Office	2014	2014/001	An evaluation of UNICEF Maldives strategies in addressing issues affecting women and children	Highly satisfactory

No.	Country	Region	Year	Sequence #	Title	Rating
30	Mali	West and Central Africa Regional Office	2014	2014/002	Évaluation du Programme WASH à l'école au Mali	Highly satisfactory
31	Mali	West and Central Africa Regional Office	2014	2014/001	Summative External Evaluation of the Catalytic Initiative (CI)/Integrated Health Systems Strengthening (IHSS) Programme in Mali	Highly satisfactory
32	Namibia	Eastern and Southern Africa Regional Office	2014	2014/004	School Based HIV Testing and Counselling Pilot Programme Evaluation	Highly satisfactory
33	Nepal	South Asia Regional Office	2015	2014/013	Evaluation of Let Us Learn Nepal: After-School Programme for Girls and Girls Access to Education Programme	Highly satisfactory
34	Pakistan	South Asia Regional Office	2014	2014/001	End of Project Evaluation for Norway-Pakistan Partnership Initiative	Highly satisfactory
35	Pakistan	South Asia Regional Office	2014	2014/003	Evaluation of the UNICEF Sanitation Programme at Scale in Pakistan (SPSP) – Phase 1 (2013-14)	Highly satisfactory
36	Pakistan	South Asia Regional Office	2014	2014/002	Evaluation of Young Champions Initiative for Girls' Education	Highly satisfactory
37	Peru	Latin America and Caribbean Regional Office	2015	2014-015	Evaluación de Medio Termino: "Mejorando la Educación Básica de Niñas y Niños en la Amazonía y el Sur Andino del Perú, 2010-2017"	Highly satisfactory
38	Philippines	East Asia and the Pacific Regional Office	2014	2014/007	IASC Inter-agency Humanitarian Evaluation of the Typhoon Haiyan Response	Highly satisfactory
39	Philippines	East Asia and the Pacific Regional Office	2014	2014/04	Real-Time Evaluation of UNICEF's Response to the Typhoon Haiyan in the Philippines	Highly satisfactory
40	Republic of Kyrgyzstan	Central & Eastern Europe, Commonwealth of Independent States RO	2014	2014/004	Evaluation of DFID/UNICEF Equity Programme	Highly satisfactory

No.	Country	Region	Year	Sequence #	Title	Rating
41	Republic of Kyrgyzstan	Central & Eastern Europe, Commonwealth of Independent States RO	2014	2014/002	Evaluation of Perinatal Care Programme	Highly satisfactory
42	Republic of Montenegro	Central & Eastern Europe, Commonwealth of Independent States RO	2014	2014/005	Evaluation Report for the Justice for Children Initiative	Highly satisfactory
43	Romania	Central & Eastern Europe, Commonwealth of Independent States RO	2014	2014/001	National Intermediate Evaluation of the "School Attendance Initiative" Model	Highly satisfactory
44	Tajikistan	Central & Eastern Europe, Commonwealth of Independent States RO	2014	2014/004	YOUTH FRIENDLY HEALTH SERVICES PROGRAM IN TAJIKISTAN, 2006-2013: Program Evaluation Report	Highly satisfactory
45	Ukraine	Central & Eastern Europe, Commonwealth of Independent States RO	2014	2014/001	Prevention of Mother-to-Child Transmission and Improving Neonatal Outcomes among Drug-dependent Pregnant Women and Children Born to Them in Ukraine	Highly satisfactory
46	United Rep. of Tanzania	Eastern and Southern Africa Regional Office	2014	2014/010	Formative Evaluation of the Children's Agenda	Highly satisfactory
47	USA	Corporate (HQ)	2014	2014/012	Formative Evaluation of UNICEF's MoRES	Highly satisfactory
48	USA	Corporate (HQ)	2013	2014/006A	UNICEF Evaluation of the Multiple Indicator Cluster Surveys (MICS) - Round 4, Evaluation Part 1: Response to lessons learned in prior rounds and preparations for Round 5	Highly satisfactory

No.	Country	Region	Year	Sequence #	Title	Rating
49	USA	Corporate (HQ)	2014	2014/006B	UNICEF Evaluation of the Multiple Indicator Cluster Surveys (MICS) - Round 4, Evaluation Part 2: MICS funding, stakeholder needs and demands and use	Highly satisfactory
50	USA	Corporate (HQ)	2014	2014/005	UNICEF's Upstream Work in Basic Education and Gender Equality 2003-2012: SYNTHESIS REPORT	Highly satisfactory
51	Zimbabwe	Corporate (HQ)	2014	2014/002	Country Case Study: Zimbabwe (Unicef's Upstream Work in Basic Education and Gender Equality 2003-2012)	Highly satisfactory
52	Barbados	Latin America and Caribbean Regional Office	2014	2014/001	Final Report for the Formative Evaluation of the HighScope Curriculum Reform Program	Mostly Satisfactory
53	Benin	West and Central Africa Regional Office	2014	2014/002	Évaluation des Quatre Innovations EDUCOM au Bénin	Mostly Satisfactory
54	Bhutan	South Asia Regional Office	2014	2104/002	Evaluation of the Weekly Iron and Folic Acid Supplementation (WFIS) Program 2004-2014, Bhutan	Mostly Satisfactory
55	Madagascar	Eastern and Southern Africa Regional Office	2014	2014/005	Evaluation de l'approche "assainissement total piloté par la communauté" (atpc)	Mostly Satisfactory
56	Malawi	Eastern and Southern Africa Regional Office	2014	2014/005	Summative report on the external evaluation of the Catalytic Initiative (CI)/ Integrated Health Systems Strengthening (IHSS) programme in Malawi	Mostly Satisfactory
57	Mali	West and Central Africa Regional Office	2014	2014/003	Final report: Impact evaluation of community-led total sanitation (CLTS) in rural Mali	Mostly Satisfactory
58	Mongolia	East Asia and the Pacific Regional Office	2014	2014/006	REDS Strategy Evaluation Mongolia	Mostly Satisfactory
59	Nigeria	West and Central Africa Regional Office	2014	2014/012	Impact Evaluation of Water, Sanitation and Hygiene (WASH) within the UNICEF Country Programme of Cooperation, Government of Nigeria and UNICEF, 2009-2013	Mostly Satisfactory
60	Rwanda	Eastern and Southern Africa Regional Office	2014	2014/001	Emergency Preparedness for the influx of refugees into Rwanda	Mostly Satisfactory

No.	Country	Region	Year	Sequence #	Title	Rating
61	Sierra Leone	West and Central Africa Regional Office	2014	2014/002	Evaluation of Journalists Training on Ethical Reporting on Child Rights Issues	Mostly Satisfactory
62	Somalia	Eastern and Southern Africa Regional Office	2014	2014/012	Regional Supply Hub Mechanism as a Strategy for Wash Emergency Response in Somalia	Mostly Satisfactory
63	Sudan	Middle East and North Africa Regional Office	2014	2014/004	Child Friendly Community Initiative - Evaluation Report, Sudan	Mostly Satisfactory
64	Sudan	Middle East and North Africa Regional Office	2014	2014/007	EVALUATION OF UNICEF SUDAN COUNTRY OFFICE FIELD DELIVERY STRUCTURE	Mostly Satisfactory
65	Tajikistan	Central & Eastern Europe, Commonwealth of Independent States RO	2014	2014/010	Juvenile Justice Alternative Project (2010-2014) - Evaluation Report	Mostly Satisfactory
66	Togo	West and Central Africa Regional Office	2014	2014/001	Évaluation des Interventions à Base Communautaire (Nutrition & ATPC) dans les Régions des Savanes et de la Kara	Mostly Satisfactory
67	Uganda	Eastern and Southern Africa Regional Office	2014	2014/001	UNICEF – Ugandan Ministry of Education and Sports BRMS Mentorship Project Evaluation	Mostly Satisfactory
69	Bhutan	South Asia Regional Office	2014	2014/001	Evaluation of WASH in schools, Bhutan	Unsatisfactory
69	Sierra Leone	West and Central Africa Regional Office	2014	2014/003	An Evaluation of the Impact of UNICEF Radio Listener Groups Project in Sierra Leone	Unsatisfactory

Appendix VI GEROS Assessment Tool

UNICEF Global Evaluation Report Oversight System (GEROS) Review Template

Colour Coding	CC	Dark green	Green	Amber	Red	White
	Questions	Outstanding	Yes	Mostly Satisfactory	No	Not Applicable
	Section & Overall Rating	Outstanding, best practice	Highly Satisfactory	Mostly Satisfactory	Unsatisfactory	

The key questions are highlighted as shown here, and are important questions in guiding the analysis of the section

The Cornerstone questions are in column J and are questions that need to be answered for rating and justification of each of the six sections

UNEG Standards for Evaluation in the UN System	UNEG Norms for Evaluation in the UN System	UNICEF Adapted UNEG Evaluation Report Standards

	Response					If the report is not an Evaluation please do not continue your review.
Title of the Evaluation Report						
Report sequence number		Date of Review		Year of the Evaluation Report		
Region				Country		
Type of Report				TORs Present		
Name of reviewer						

Classification of Evaluation Report		Comments		
Geographic Scope *(Coverage of the programme being evaluated & generalizability of evaluation findings)*				
Management of Evaluation *(Managerial control and oversight of evaluation decisions)*				
Purpose *(Speaks to the overarching goal for conducting the evaluation; its raison d'être)*				

Result *(Level of changes sought, as defined in RBM: refer to substantial use of highest level reached)*					
SPOA Correspondence *(Alignment with SPOA focus area priorities: (1) Young child survival and development; (2) Basic education and gender equality; (3) HIV/AIDS and children; (4) Child protection from violence, exploitation and abuse; and (5) Policy advocacy and partnerships for children's rights)*					
Level of Independence *(Implementation and control of the evaluation activities)*					
Approach					
SECTION A: OBJECT OF THE EVALUATION				Second Review (dd/mm/yy)	Third Review (dd/mm/yy)
Question	cc	Remarks	A/ Does the report present a clear & full description of the 'object' of the evaluation? The report should describe the object of the evaluation including the results chain, meaning the 'theory of change' that underlies the programme being evaluated. This theory of change includes what the programme was meant to achieve and the pathway (chain of results) through which it was expected to achieve this. The context of key social, political, economic,	Constructive feedback for future reports *Including how to address weaknesses and maintaining good practice*	A/ Does the report present a clear & full description of the 'object' of the evaluation?	A/ Does the report present a clear & full description of the 'object' of the evaluation?
Object and context					
1 Is the object of the evaluation well described? This needs to include a clear description of the interventions (project, programme, policies, otherwise) to be evaluated including how the designer thought that it would address the problem identified, implementing modalities, other parameters including costs, relative importance in the organization and (number of) people reached.					
2 Is the context explained and related to the object that is to be evaluated? The context includes factors that have a direct bearing on the object of the evaluation: social, political, economic, demographic, institutional. These factors may include strategies, policies, goals, frameworks & priorities at the: international level; national Government level; individual agency level					
3 Does this illuminate findings? The context should ideally be linked to the findings so that it is clear how the wider situation may have influenced the outcomes observed.					

			demographic, and institutional factors that have a direct bearing on the object should be described. For example, the partner government's strategies and priorities, international, regional or country development goals, strategies and frameworks, the concerned agency's corporate goals & priorities, as appropriate.							
Theory of Change										
4 Is the results chain or logic well articulated? The report should identify how the designers of the evaluated object thought that it would address the problem that they had identified. This can include a results chain or other logic models such as theory of change. It can include inputs, outputs and outcomes, it may also include impacts. The models need to be clearly described and explained.										
Stakeholders and their contributions										
5 Are key stakeholders clearly identified? These include o implementing agency(ies) o development partners o rights holders o primary duty bearers o secondary duty bearers										
6 Are key stakeholders' contributions described? This can involve financial or other contributions and should be specific. If joint program also specify UNICEF contribution, but if basket funding question is not applicable										
7 Are UNICEF contributions described? This can involve financial or other contributions and should be specific										
Implementation Status										
8 Is the implementation status described? This includes the phase of implementation and significant changes that have happened to plans, strategies, performance frameworks, etc that have occurred - including the implications of these										

changes									
Executive Feedback on Section A Issues for this section relevant for feedback to senior management (positives & negatives), & justify rating. *Up to two sentences*									
SECTION B: EVALUATION PURPOSE, OBJECTIVES AND SCOPE						Second Review		Third Review	
Question	cc	Remarks	B/ Are the evaluation's purpose, objectives and scope sufficiently clear to guide the evaluation? The purpose of the evaluation should be clearly defined, including why the evaluation was needed at that point in time, who needed the information, what information is needed, and how the information will be used. The report should provide a clear explanation of the evaluation objectives and scope including main evaluation questions and describes and justifies what the evaluation did and did not cover. The report should describe and provide an explanation of the chosen evaluation criteria, performance standards, or other	Constructive feedback for future reports *Including how to address weaknesses and maintaining good practice*	B/ Are the evaluation's purpose, objectives and scope sufficiently clear to guide the evaluation?		B/ Are the evaluation's purpose, objectives and scope sufficiently clear to guide the evaluation?	
Purpose, objectives and scope									
9 Is the purpose of the evaluation clear? This includes why the evaluation is needed at this time, who needs the information, what information is needed, how the information will be used.									
10 Are the objectives and scope of the evaluation clear and realistic? This includes: Objectives should be clear and explain what the evaluation is seeking to achieve; Scope should clearly describe and justify what the evaluation will and will not cover; Evaluation questions may optionally be included to add additional details									
11 Do the objective and scope relate to the purpose? The reasons for holding the evaluation at this time in the project cycle (purpose) should link logically with the specific objectives the evaluation seeks to achieve and the boundaries chosen for the evaluation (scope)									
Evaluation framework									

			criteria used by the evaluators.						
12 Does the evaluation provide a relevant list of evaluation criteria that are explicitly justified as appropriate for the Purpose? It is imperative to make the basis of the value judgements used in the evaluation transparent if it is to be understood and convincing. UNEG evaluation standards refer to the OECD/DAC criteria, but other criteria can be used such as Human rights and humanitarian criteria and standards (e.g. SPHERE Standards) but this needs justification.. Not all OECD/DAC criteria are relevant to all evaluation objectives and scopes. The TOR may set the criteria to be used, but these should be (re)confirmed by the evaluator. Standard OECD DAC Criteria include: Relevance; Effectiveness; Efficiency; Sustainability; Impact Additional humanitarian criteria include; Coverage; Coordination; Coherence; Protection; timeliness; connectedness; appropriateness. *(This is an extremely important question to UNICEF)*									
13 Does the evaluation explain why the evaluation criteria were chosen and/or any standard DAC evaluation criteria (above) rejected? The rationale for using each particular non-OECD-DAC criterion (if applicable) and/or rejecting any standard OECD-DAC criteria (where they would be applicable) should be explained in the report.									
Executive Feedback on Section B Issues for this section relevant for feedback to senior management (positives & negatives), & justify rating. *Up to two sentences*									
SECTION C: EVALUATION METHODOLOGY, GENDER, HUMAN RIGHTS AND EQUITY						Second Review		Third Review	
Question	cc	Remarks	C/ Is the methodology appropriate and sound? The report should present a transparent description of the methodology applied to the evaluation that clearly explains how the evaluation was specifically designed to		Constructive feedback for future reports *Including how to address weaknesses and maintaining good practice*	C/ Is the methodology appropriate and sound?		C/ Is the methodology appropriate and sound?	
Data collection									
14 Does the report specify data collection methods, analysis methods, sampling methods and benchmarks? This should include the rationale for selecting methods and their limitations based on commonly accepted best practice.									
15 Does the report specify data sources, the rationale for their selection, and their limitations? This should include a discussion of how the mix of data sources was used to obtain a diversity of perspectives, ensure accuracy & overcome data limits									
Ethics									

16 Are ethical issues and considerations described? The design of the evaluation should contemplate: How ethical the initial design of the programme was; The balance of costs and benefits to participants (including possible negative impact) in the programme and in the evaluation; The ethics of who is included and excluded in the evaluation and how this is done			address the evaluation criteria, yield answers to the evaluation questions and achieve the evaluation purposes. The report should also present a sufficiently detailed description of methodology in which methodological choices are made explicit and justified and in which limitations of methodology applied are included. The report should give the elements to assess the appropriateness of the methodology. Methods as such are not 'good' or 'bad', they are only so in relation to what one tries to get to know as part of an evaluation. Thus this standard assesses the suitability of the methods selected for the specifics of the evaluation concerned, assessing if the methodology is suitable to the subject matter and the information collected are sufficient to meet the evaluation			
17 Does the report refer to ethical safeguards appropriate for the issues described? When the topic of an evaluation is contentious, there is a heightened need to protect those participating. These should be guided by the UNICEF Evaluation Office Technical Note and include: protection of confidentiality; protection of rights; protection of dignity and welfare of people (especially children); Informed consent; Feedback to participants; Mechanisms for shaping the behaviour of evaluators and data collectors						

			objectives.						
Results Based Management									
18 Is the capability and robustness of the evaluated object's monitoring system adequately assessed? The evaluation should consider the details and overall functioning of the management system in relation to results: from the M&E system design, through individual tools, to the use of data in management decision making.									
19 Does the evaluation make appropriate use of the M&E framework of the evaluated object? In addition to articulating the logic model (results chain) used by the programme, the evaluation should make use of the object's logframe or other results framework to guide the assessment. The results framework indicates how the programme design team expected to assess effectiveness, and it forms the guiding structure for the management of implementation.									
Human Rights, Gender and Equity									
20 Did the evaluation design and style consider incorporation of the UN and UNICEF's commitment to a human rights-based approach to programming, to gender equality, and to equity? This could be done in a variety of ways including: use of a rights-based framework, use of CRC, CCC, CEDAW and other rights related benchmarks, analysis of right holders and duty bearers and focus on aspects of equity, social exclusion and gender. Style includes: using human-rights language; gender-sensitive and child-sensitive writing; disaggregating data by gender, age and disability groups; disaggregating data by socially excluded groups. Promote gender-sensitive interventions as a core programmatic priority, To the extent possible, all relevant policies, programmes and activities will mainstream gender equality.									
21 Does the evaluation assess the extent to which the implementation of the evaluated object was monitored through human rights (inc. gender, equity & child rights) frameworks? UNICEF commits to go beyond monitoring the achievement of desirable outcomes, and to ensure that these are achieved through morally acceptable processes. The evaluation should consider whether the programme was managed and adjusted according to human rights and gender monitoring of processes.									
22 Do the methodology, analytical framework, findings, conclusions, recommendations & lessons provide appropriate information on HUMAN RIGHTS (inc. women & child rights)? The inclusion of human rights frameworks in the evaluation methodology should continue to cascade down the evaluation report and be obvious in the data analysis, findings, conclusions, any recommendations and any lessons learned. If identified in the									

scope the methodology should be capable of assessing the level of: Identification of the human rights claims of rights-holders and the corresponding human rights obligations of duty-bearers, as well as the immediate underlying & structural causes of the non realisation of rights.; Capacity development of rights-holders to claim rights, and duty-bearers to fulfil obligations. Support for humanitarian action – achieving faster scaling up of response, early identification of priorities and strategies, rapid deployment of qualified staff and clear accountabilities and responses consistent with humanitarian principles in situations of unrest or armed conflict.									
23 Do the methodology, analytical framework, findings, conclusions, recommendations & lessons provide appropriate information on GENDER EQUALITY AND WOMEN'S EMPOWERMENT? The inclusion of gender equality frameworks in the evaluation methodology should continue to cascade down the evaluation report and be obvious in the data analysis, findings, conclusions, any recommendations and any lessons learned. If identified in the scope the methodology should be capable of assessing the immediate underlying & structural causes of social exclusion; and capacity development of women to claim rights, and duty-bearers to fulfil their equality obligations.									
24 Do the methodology, analytical framework, findings, conclusions, recommendations & lessons provide appropriate information on EQUITY? The inclusion of equity considerations in the evaluation methodology should continue to cascade down the evaluation report and be obvious in the data analysis, findings, conclusions, any recommendations and any lessons learned. If identified in the scope the methodology should be capable of assessing the capacity development of rights-holders to claim rights, and duty-bearers to fulfil obligations & aspects of equity.									
Stakeholder participation									
25 Are the levels and activities of stakeholder consultation described? This goes beyond just using stakeholders as sources of information and includes the degree of participation in the evaluation itself. The report should include the rationale for selecting this level of participation. Roles for participation might include: o Liaison o Technical advisory o Observer o Active decision making The reviewer should look for the soundness of the description and rationale for the degree of participation rather than the level of participation itself.									
26 Are the levels of participation appropriate for the task in hand? The breadth & degree of stakeholder participation feasible in evaluation activities will depend partly on the kind of									

participation achieved in the evaluated object. The reviewer should note here whether a higher degree of participation may have been feasible & preferable.									
Methodological robustness									
27 Is there an attempt to construct a counterfactual or address issues of contribution/attribution? The counterfactual can be constructed in several ways which can be more or less rigorous. It can be done by contacting eligible beneficiaries that were not reached by the programme, or a theoretical counterfactual based on historical trends, or it can also be a comparison group.									
28 Does the methodology facilitate answers to the evaluation questions in the context of the evaluation? The methodology should link back to the Purpose and be capable of providing answers to the evaluation questions.									
29 Are methodological limitations acceptable for the task in hand? Limitations must be specifically recognised and appropriate efforts taken to control bias. This includes the use of triangulation, and the use of robust data collection tools (interview protocols, observation tools etc). Bias limitations can be addressed in three main areas: Bias inherent in the sources of data; Bias introduced through the methods of data collection; Bias that colours the interpretation of findings									
Executive Feedback on Section C Issues for this section relevant for feedback to senior management (positives & negatives), & justify rating. *Up to two sentences*									
SECTION D: FINDINGS AND CONCLUSIONS						Second Review		Third Review	
Question	cc	Remarks	D/ Are the findings and conclusions, clearly presented, relevant and based on evidence & sound analysis? Findings should respond directly to the evaluation criteria and questions detailed in the scope and objectives section of the report. They should be based on	Constructive feedback for future reports *Including how to address weaknesses and maintaining good practice*		D/ Are the findings and conclusions, clearly presented, relevant and based on evidence & sound analysis?		D/ Are the findings and conclusions, clearly presented, relevant and based on evidence & sound analysis?	
Completeness and logic of findings									
30 Are findings clearly presented and based on the objective use of the reported evidence? Findings regarding the inputs for the completion of activities or process achievements should be distinguished clearly from results. Findings on results should clearly distinguish outputs, outcomes and impacts (where appropriate). Findings must demonstrate full marshalling and objective use of the evidence generated by the evaluation data collection. Findings should also tell the 'whole story' of the evidence and avoid bias.									
31 Do the findings address all of the evaluation's stated criteria and questions?									

The findings should seek to systematically address all of the evaluation questions according to the evaluation framework articulated in the report.			evidence derived from data collection and analysis methods described in the methodology section of the report. Conclusions should present reasonable judgments based on findings and substantiated by evidence, providing insights pertinent to the object and purpose of the evaluation.							
32 Do findings demonstrate the progression to results based on the evidence reported? There should be a logical chain developed by the findings, which shows the progression (or lack of) from implementation to results.										
33 Are gaps and limitations discussed? The data may be inadequate to answer all the evaluation questions as satisfactorily as intended, in this case the limitations should be clearly presented and discussed. Caveats should be included to guide the reader on how to interpret the findings. Any gaps in the programme or unintended effects should also be addressed.										
34 Are unexpected findings discussed? If the data reveals (or suggests) unusual or unexpected issues, these should be highlighted and discussed in terms of their implications.										
Cost Analysis										
35 Is a cost analysis presented that is well grounded in the findings reported? Cost analysis is not always feasible or appropriate. If this is the case then the reasons should be explained. Otherwise the evaluation should use an appropriate scope and methodology of cost analysis to answer the following questions: o How programme costs compare to other similar programmes or standards o Most efficient way to get expected results o Cost implications of scaling up or down o Cost implications for replicating in a different context o Is the programme worth doing from a cost perspective o Costs and the sustainability of the										

programme.									
Contribution and causality									
36 Does the evaluation make a fair and reasonable attempt to assign contribution for results to identified stakeholders? For results attributed to the programme, the result should be mapped as accurately as possible to the inputs of different stakeholders.									
37 Are causal reasons for accomplishments and failures identified as much as possible? These should be concise and usable. They should be based on the evidence and be theoretically robust. *(This is an extremely important question to UNICEF)*									
Strengths, weaknesses and implications									
38 Are the future implications of continuing constraints discussed? The implications can be, for example, in terms of the cost of the programme, ability to deliver results, reputational risk, and breach of human rights obligations.									
39 Do the conclusions present both the strengths and weaknesses of the evaluated object? Conclusions should give a balanced view of both the stronger aspects and weaker aspects of the evaluated object with reference to the evaluation criteria and human rights based approach.									
Completeness and insight of conclusions									
40 Do the conclusions represent actual insights into important issues that add value to the findings? Conclusions should go beyond findings and identify important underlying problems and/or priority issues. Simple conclusions that are already well known do not add value and should be avoided.									
41 Do conclusions take due account of the views of a diverse cross-section of stakeholders? As well as being logically derived from findings, conclusions should seek to represent the range of views encountered in the evaluation, and not simply reflect the bias of the individual evaluator. Carrying these diverse views through to the presentation of conclusions (considered here) is only possible if the methodology has gathered and analysed information from a broad range of stakeholders.									
42 Are the conclusions pitched at a level that is relevant to the end users of the evaluation? Conclusions should speak to the evaluation participants, stakeholders and users. These may cover a wide range of groups									

and conclusions should thus be stated clearly and accessibly: adding value and understanding to the report (for example, some stakeholders may not understand the methodology or findings, but the conclusions should clarify what these findings mean to them in the context of the programme).									
Executive Feedback on Section D Issues for this section relevant for feedback to senior management (positives & negatives), & justify rating. *Up to two sentences*									
SECTION E: RECOMMENDATIONS AND LESSONS LEARNED						Second Review		Third Review	
Question	cc	Remarks	E/ Are the recommendations and lessons learned relevant and actionable? Recommendations should be relevant and actionable to the object and purpose of the evaluation, be supported by evidence and conclusions, and be developed with involvement of relevant stakeholders. Recommendations should clearly identify the target group for each recommendation, be clearly stated with priorities for action, be actionable and reflect an understanding of the commissioning organization and potential constraints to follow up.	Constructive feedback for future reports *Including how to address weaknesses and maintaining good practice*	E/ Are the recommendations and lessons learned relevant and actionable?		E/ Are the recommendations and lessons learned relevant and actionable?	
Relevance and clarity of recommendations									
43 Are the recommendations well-grounded in the evidence and conclusions reported? Recommendations should be logically based in findings and conclusions of the report.									
44 Are recommendations relevant to the object and the purpose of the evaluation? Recommendations should be relevant to the evaluated object									
45 Are recommendations clearly stated and prioritised? If the recommendations are few in number (up to 5) then this can also be considered to be prioritised. Recommendations that are over-specific or represent a long list of items are not of as much value to managers. Where there is a long list of recommendations, the most important should be ordered in priority.									
Usefulness of recommendations									

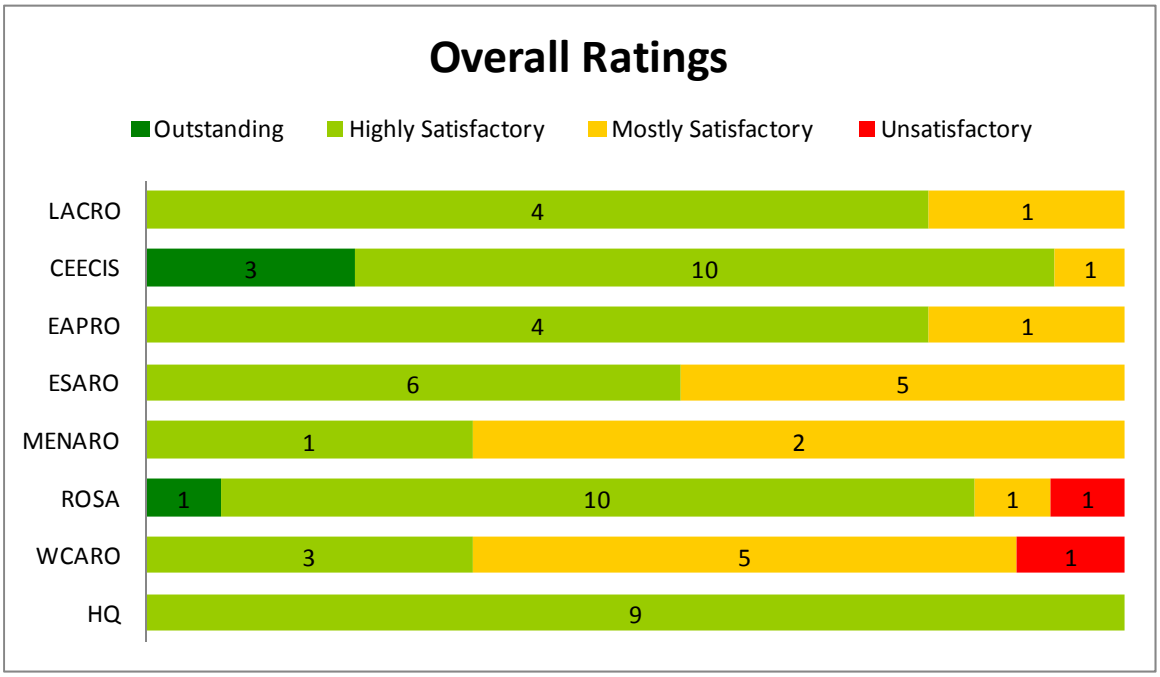
46 Does each recommendation clearly identify the target group for action? Recommendations should provide clear and relevant suggestions for action linked to the stakeholders who might put that recommendation into action. This ensures that the evaluators have a good understanding of the programme dynamics and that recommendations are realistic.									
47 Are the recommendations realistic in the context of the evaluation? This includes: o an understanding of the commissioning organisation o awareness of the implementation constraints o an understanding of the follow-up processes									
48 Does the report describe the process followed in developing the recommendations? The preparation of recommendations needs to suit the evaluation process. Participation by stakeholders in the development of recommendations is strongly encouraged to increase ownership and utility.									
Appropriate lessons learned									
49 Are lessons learned correctly identified? Lessons learned are contributions to general knowledge. They may refine or add to commonly accepted understanding, but should not be merely a repetition of common knowledge. Findings and conclusions specific to the evaluated object are not lessons learned.									
50 Are lessons learned generalised to indicate what wider relevance they may have? Correctly identified lessons learned should include an analysis of how they can be applied to contexts and situations outside of the evaluated object.									
Executive Feedback on Section E Issues for this section relevant for feedback to senior management (positives & negatives), & justify rating. *Up to two sentences*									
SECTION F: REPORT IS WELL STRUCTURED, LOGIC AND CLEAR						Second Review		Third Review	
Question	cc	Remarks	F/ Overall, do all these elements come together in a well structured, logical, clear and complete report? The report should		Constructive feedback for future reports *Including how to address*	F/ Overall, do all these elements come together in a well structured, logical, clear and complete report?		F/ Overall, do all these elements come together in a well structured, logical, clear and complete report?	
Style and presentation									
51. Do the opening pages contain all the basic elements? Basic elements include all of: Name of the evaluated object; Timeframe of the evaluation and date of the report; Locations of the evaluated object; Names and/or organisations of evaluators;									

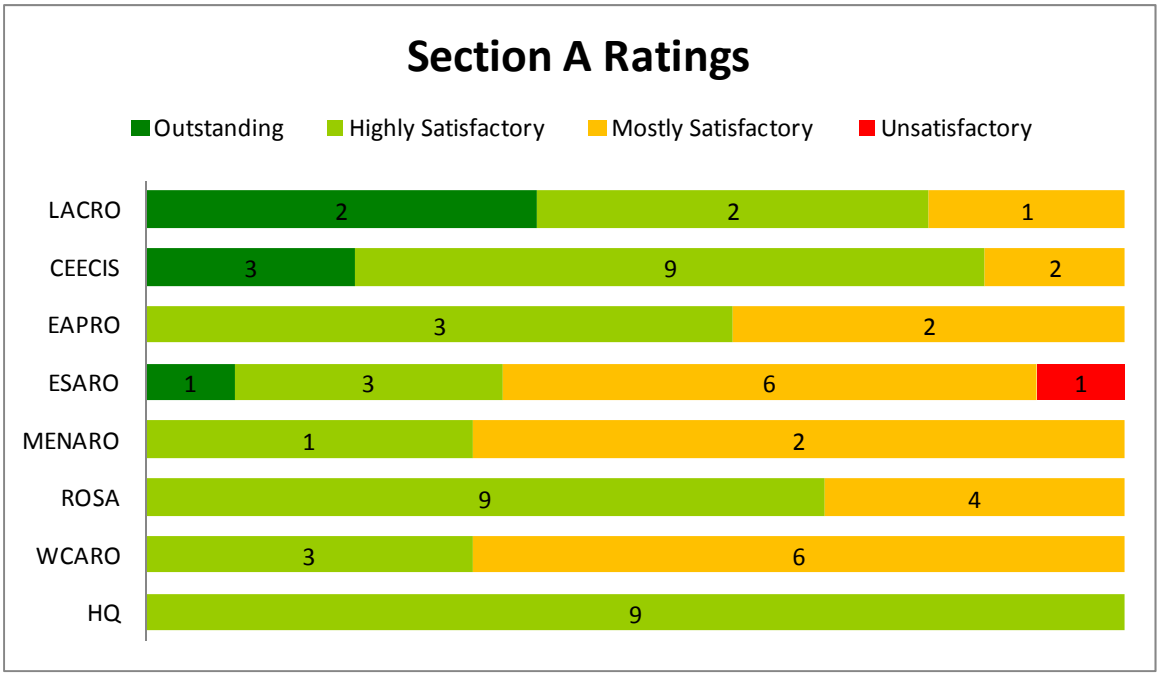
Name of the organisation commissioning the evaluation; Table of contents including tables, graphs, figures and annex; List of acronyms			be logically structured with clarity and coherence (e.g. background and objectives are presented before findings, and findings are presented before conclusions and recommendations) . It should read well and be focused.		*weaknesses and maintaining good practice*				
52 Is the report logically structured? Context, purpose, methodology and findings logically structured. Findings would normally come before conclusions, recommendations & lessons learnt									
53 Do the annexes contain appropriate elements? Appropriate elements may include: ToRs; List of interviewees and site visits; List of documentary evidence; Details on methodology; Data collection instruments; Information about the evaluators; Copy of the evaluation matrix; Copy of the Results chain. Where they add value to the report									
54 Do the annexes increase the usefulness and credibility of the report?									
Executive Summary									
55. Is an executive summary included as part of the report? If the answer is No, question 56 to 58 should be N/A									
56 Does the executive summary contain all the necessary elements? Necessary elements include all of: Overview of the evaluated object; Evaluation objectives and intended audience; Evaluation methodology; Most important findings and conclusions; Main recommendations									
57 Can the executive summary stand alone? It should not require reference to the rest of the report documents and should not introduce new information or arguments									
58 Can the executive summary inform decision making? It should be short (ideally 2-3 pages), and increase the utility for decision makers by highlight key priorities.									

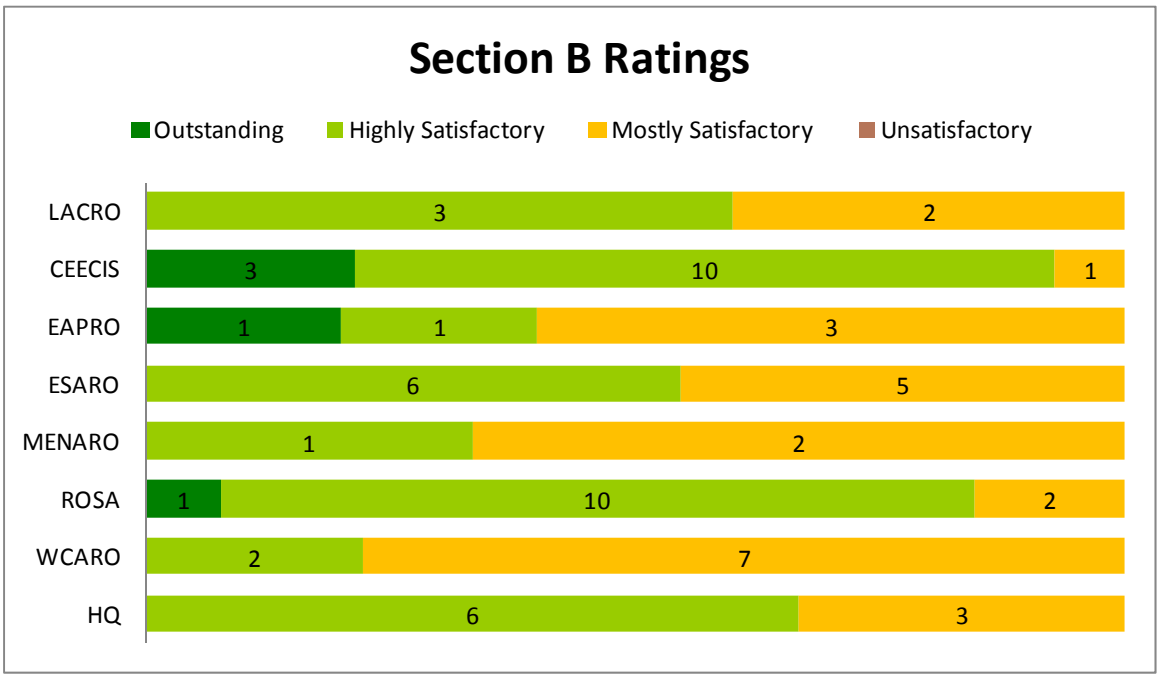
Executive Feedback on Section F Issues for this section relevant for feedback to senior management (positives & negatives), & justify rating. *Up to two sentences*							
Additional Information					Second Review		Third Review
Question	Remarks				Remarks		Remarks
i/ Does the evaluation successfully address the Terms of Reference? If the report does not include a TOR then a recommendation should be given to ensure that all evaluations include the TOR in the future. Some evaluations may be flawed because the TORs are inappropriate, too little time etc. Or, they may succeed despite inadequate TORs. This should be noted under vii in the next section							
ii/ Identify aspects of good practice in the evaluation In terms of evaluation							
iii/ Identify aspects of good practice of the evaluation In terms of programmatic, sector specific, thematic expertise							
OVERALL RATING					Second Review		Third Review
Question	cc	Remarks	OVERALL RATING Informed by the answers above, apply the reasonable person test to answer the following question: Ω/ Is this a credible report that addresses the evaluation purpose and objectives based on evidence, and that can therefore be used with confidence? This question should be considered from the perspective of UNICEF strategic management.		OVERALL RATING Informed by the answers above, apply the reasonable person test to answer the following question: Ω/ Is this a credible report that addresses the evaluation purpose and objectives based on evidence, and that can therefore be used with confidence?		OVERALL RATING Informed by the answers above, apply the reasonable person test to answer the following question: Ω/ Is this a credible report that addresses the evaluation purpose and objectives based on evidence, and that can therefore be used with confidence?
i/ To what extent does each of the six sections of the evaluation provide sufficient credibility to give the reasonable person confidence to act? Taken on their own, could a reasonable person have confidence in each of the five core evaluation elements separately? It is particularly important to consider: o Is the report methodologically appropriate? o Is the evidence sufficient, robust and authoritative? o Do the analysis, findings, conclusions and recommendations hold together?							

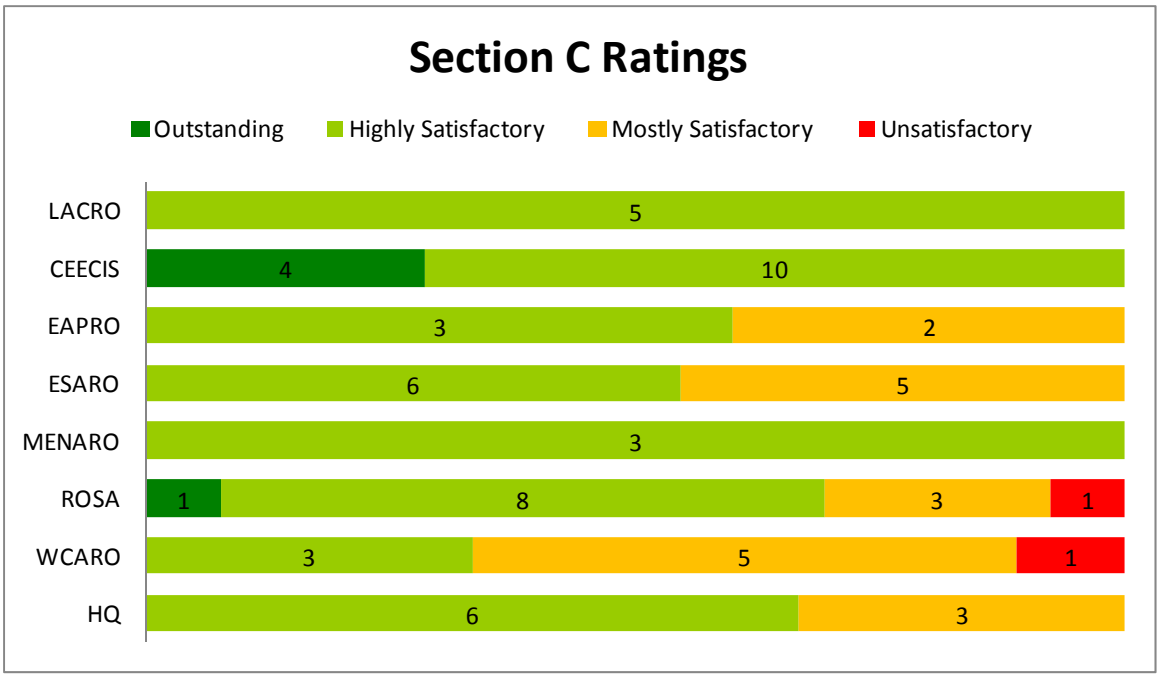
ii/ To what extent do the six sections hold together in a logically consistent way that provides common threads throughout the report? The report should hold together not just as individually appropriately elements, but as a consistent and logical 'whole'.								
iii/ Are there any reasons of note that might explain the overall performance or particular aspects of this evaluation report? This is a chance to note mitigating factors and/or crucial issues apparent in the review of the report.	ToRs							
	Others							
Executive Feedback on Overall Rating Issues for this section relevant for feedback to senior management (positives & negatives), & justify rating. *Up to two sentences*								

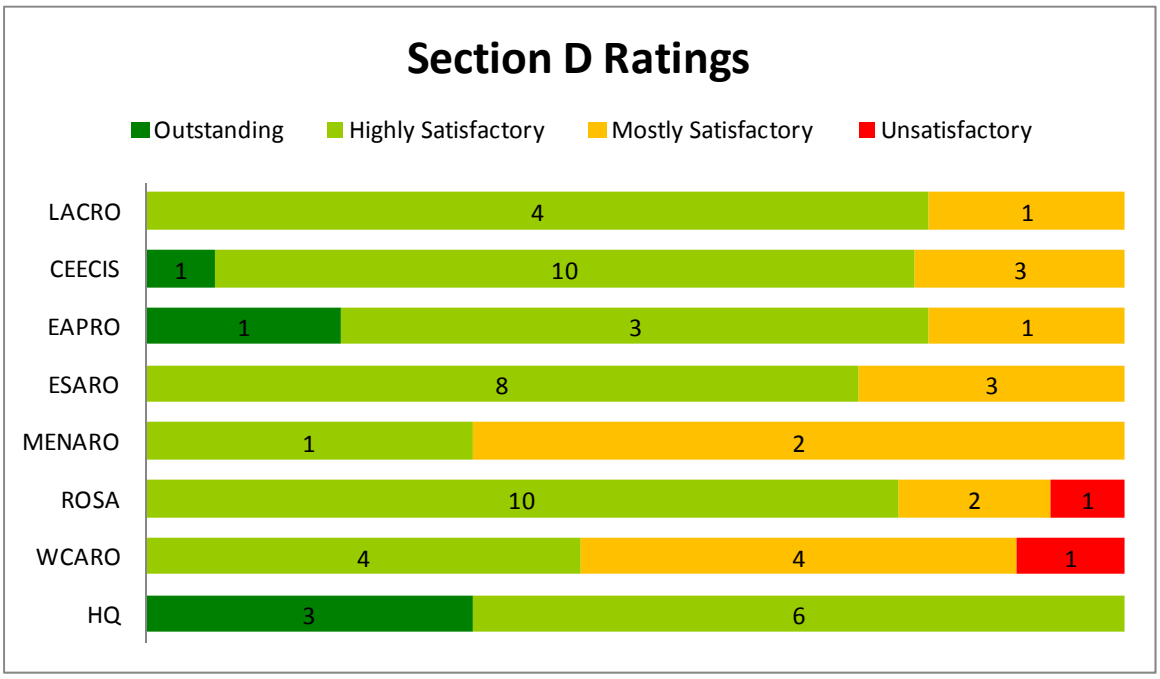
Appendix VII Regional Ratings Graphs

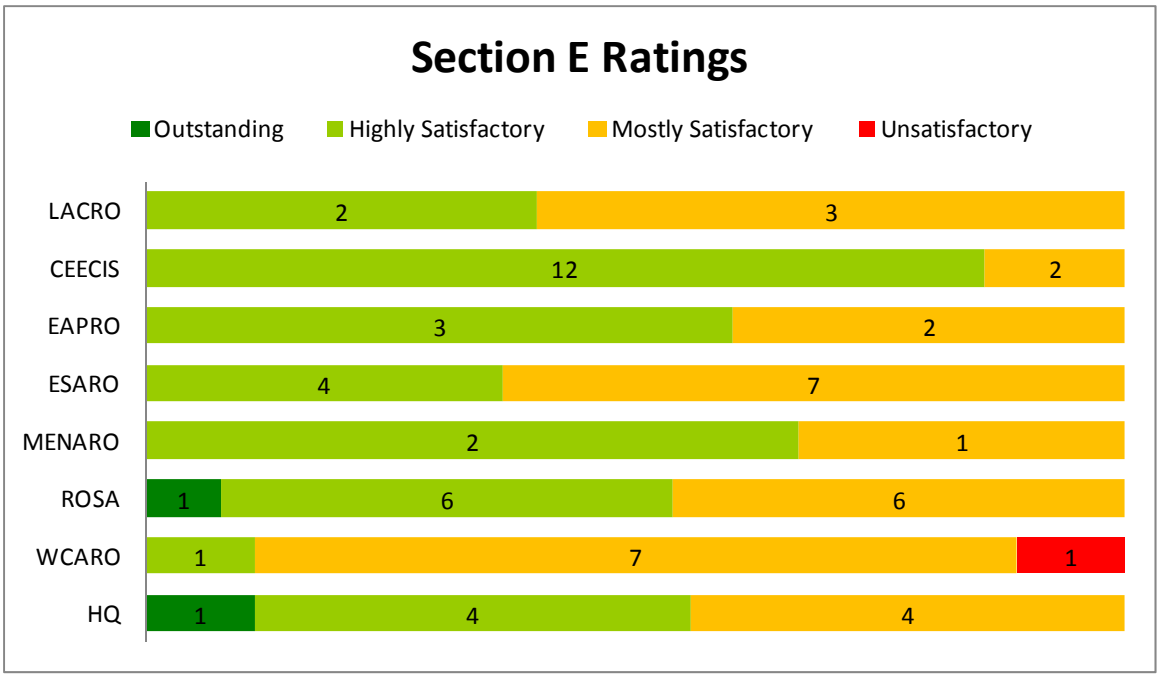


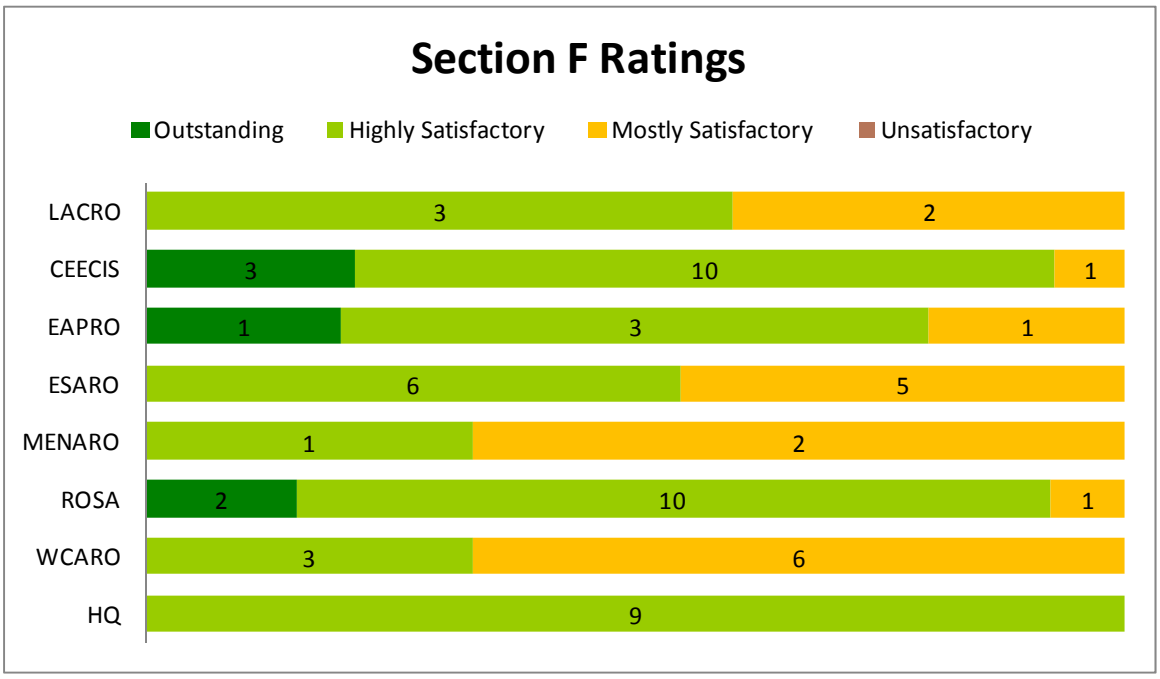



Section B Ratings
Outstanding
Highly Satisfactory
Mostly Satisfactory
Unsatisfactory
LACRO
3
2
CEECIS
3
10
1
EAPRO
1
1
3
ESARO
6
5
MENARO
1
2
ROSA
1
10
2
WCARO
2
7
HQ
6
3

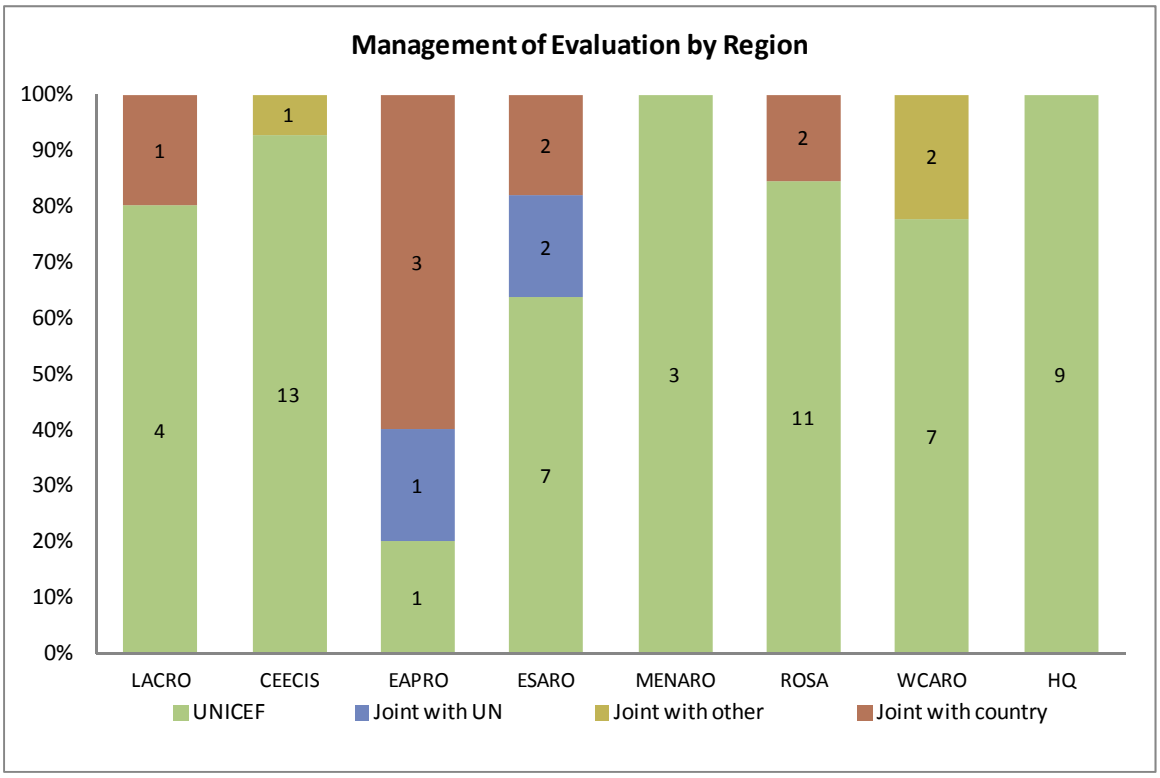

Section C Ratings
Outstanding
Highly Satisfactory
Mostly Satisfactory
Unsatisfactory
LACRO
5
CEECIS
4
10
EAPRO
3
2
ESARO
6
5
MENARO
3
ROSA
1
8
3
1
WCARO
3
5
1
HQ
6
3

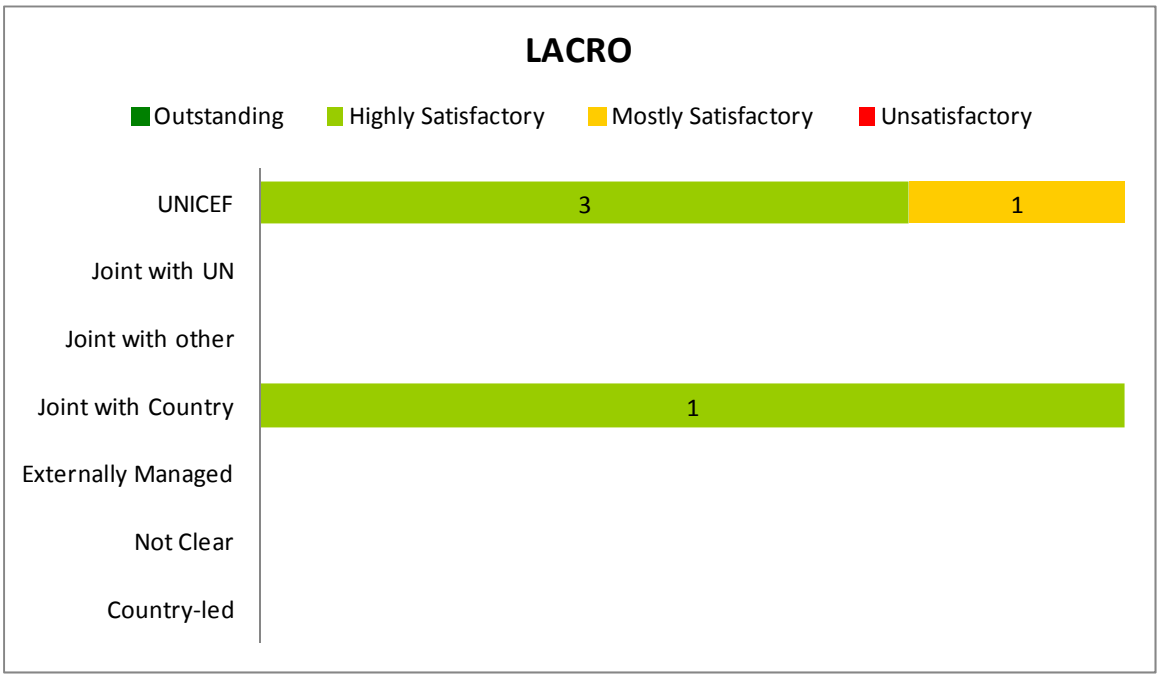

Section D Ratings
Outstanding
Highly Satisfactory
Mostly Satisfactory
Unsatisfactory
LACRO
CEECIS
EAPRO
ESARO
MENARO
ROSA
WCARO
HQ
4
1
1
10
3
1
3
1
8
3
1
2
10
2
1
4
4
1
3
6

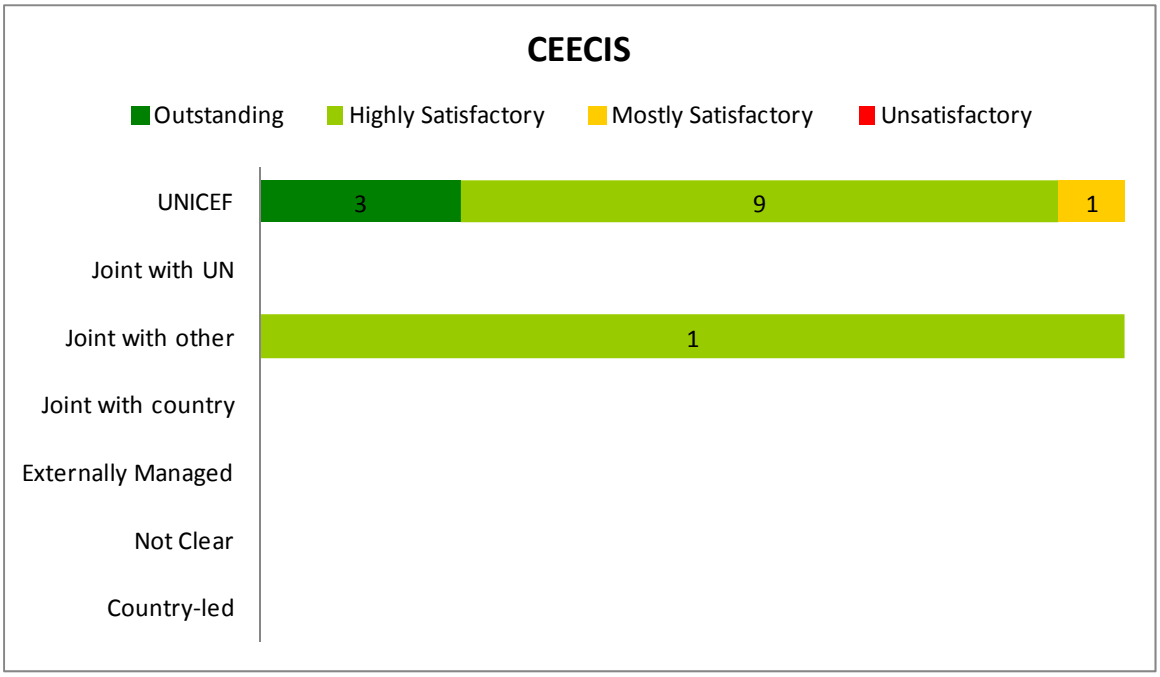

Section E Ratings
Outstanding
Highly Satisfactory
Mostly Satisfactory
Unsatisfactory
LACRO
CEECIS
EAPRO
ESARO
MENARO
ROSA
WCARO
HQ
2
3
12
2
3
2
4
7
2
1
1
6
6
1
7
1
1
4
4

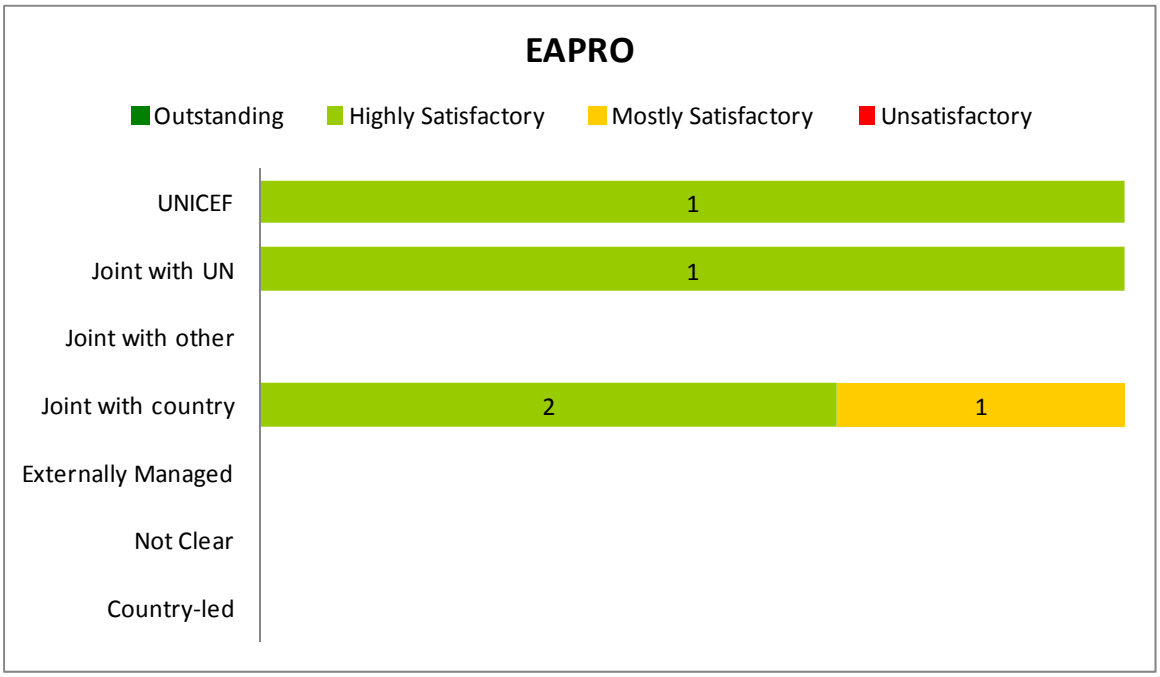

Section F Ratings
Outstanding
Highly Satisfactory
Mostly Satisfactory
Unsatisfactory
LACRO
3
2
CEECIS
3
10
1
EAPRO
1
3
1
ESARO
6
5
MENARO
1
2
ROSA
2
10
1
WCARO
3
6
HQ
9

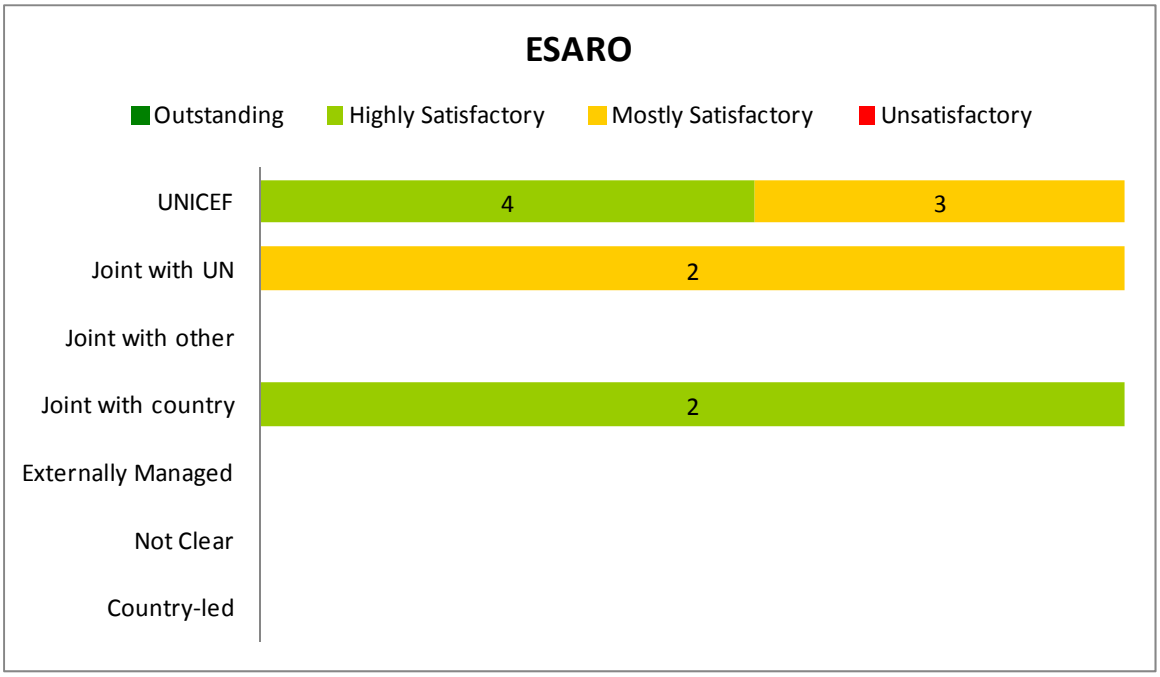
Appendix VIII UNICEF, Country-led and Jointly Managed Evaluations by Region

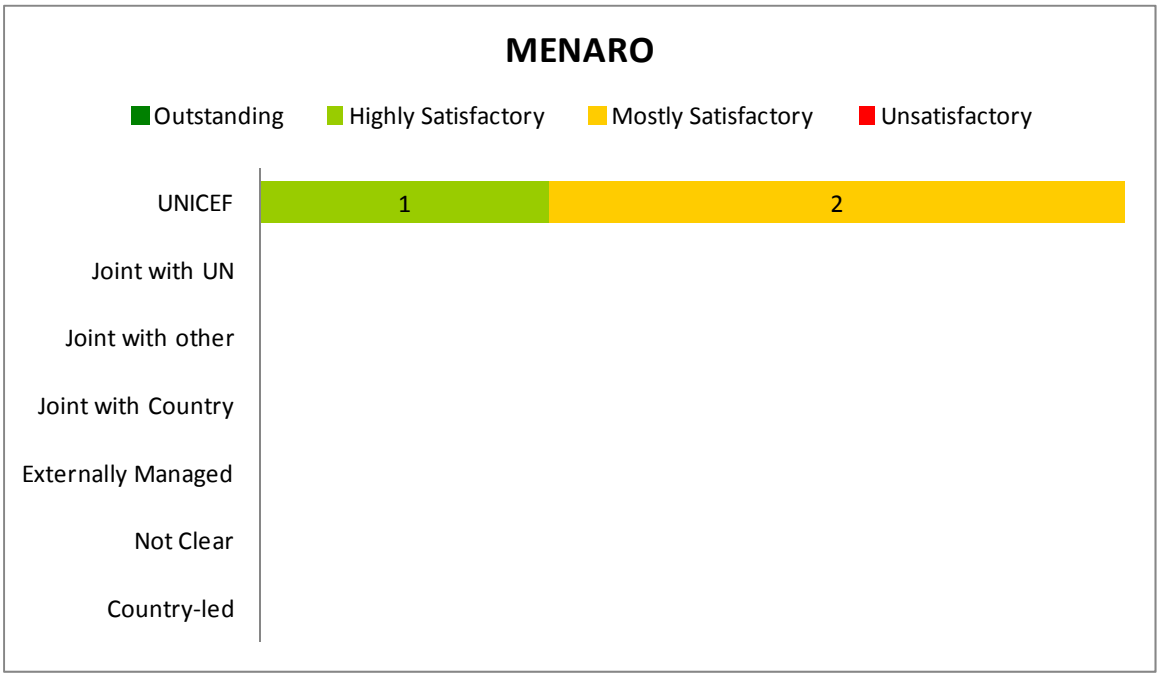


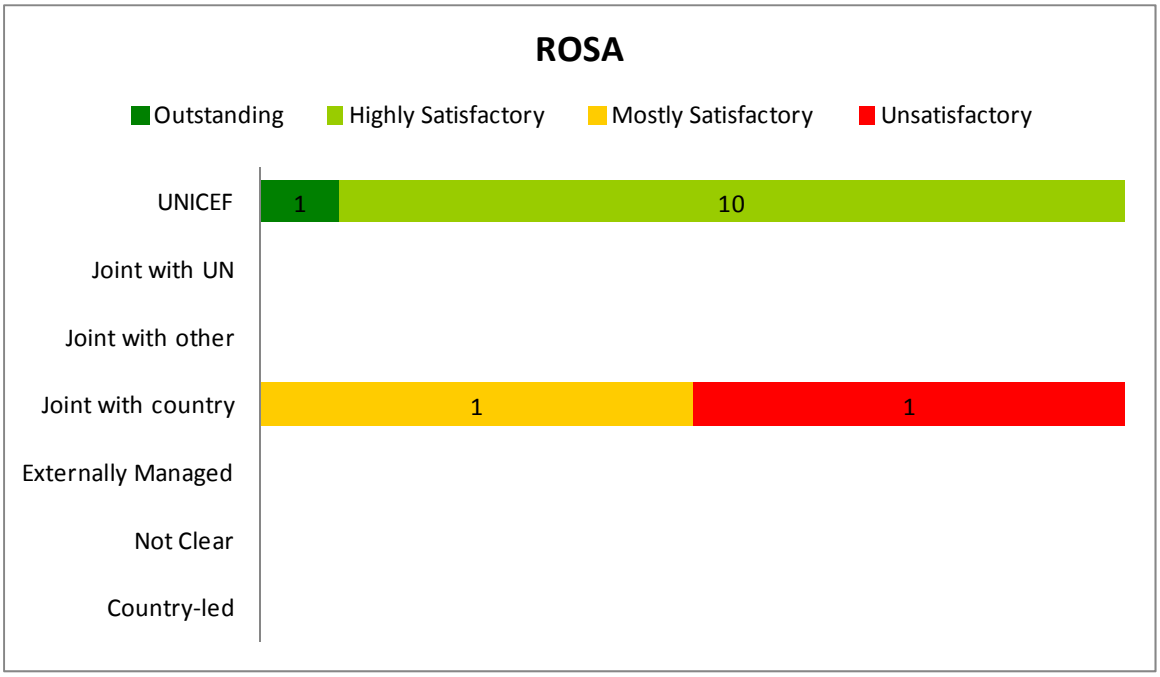


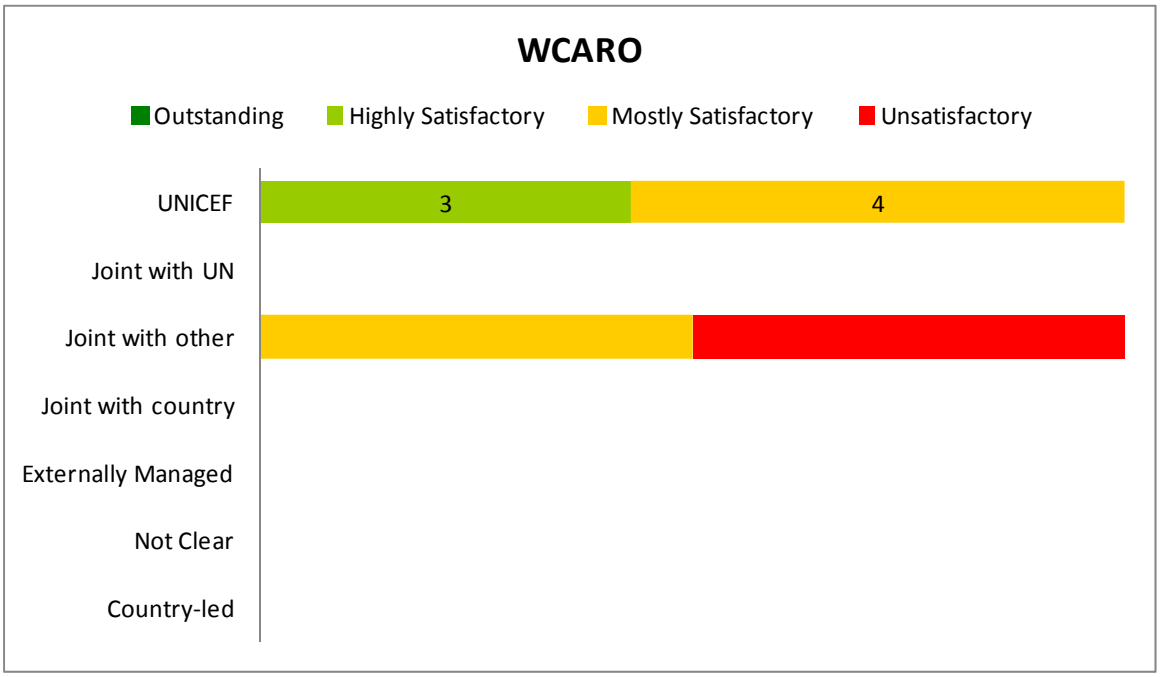

CEECIS
Outstanding
Highly Satisfactory
Mostly Satisfactory
Unsatisfactory
UNICEF
Joint with UN
Joint with other
Joint with country
Externally Managed
Not Clear
Country-led
3
9
1
1

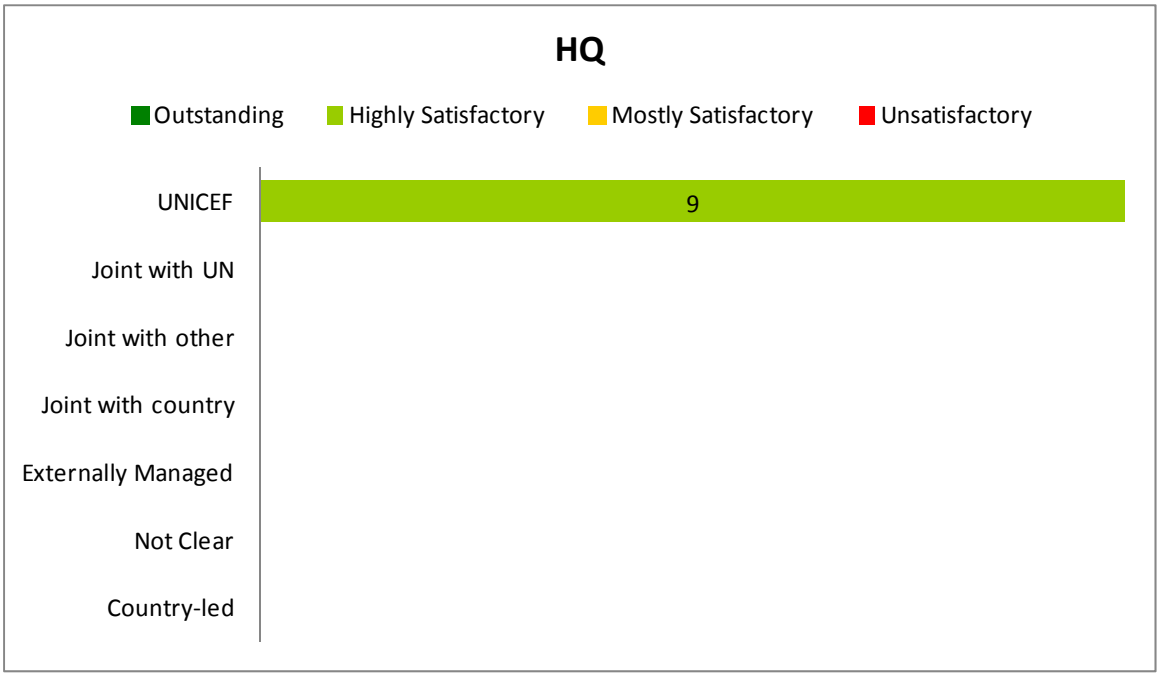

EAPRO
Outstanding
Highly Satisfactory
Mostly Satisfactory
Unsatisfactory
UNICEF
Joint with UN
Joint with other
Joint with country
Externally Managed
Not Clear
Country-led
1
1
2
1


ESARO
Outstanding
Highly Satisfactory
Mostly Satisfactory
Unsatisfactory
UNICEF
Joint with UN
Joint with other
Joint with country
Externally Managed
Not Clear
Country-led
4
3
2
2

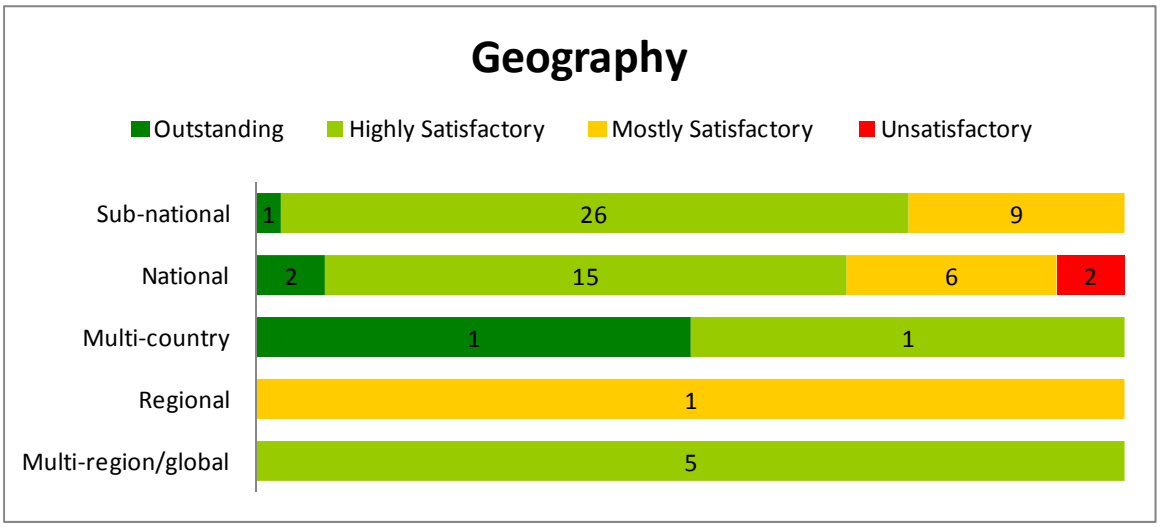

MENARO
Outstanding
Highly Satisfactory
Mostly Satisfactory
Unsatisfactory
UNICEF
Joint with UN
Joint with other
Joint with Country
Externally Managed
Not Clear
Country-led
1
2

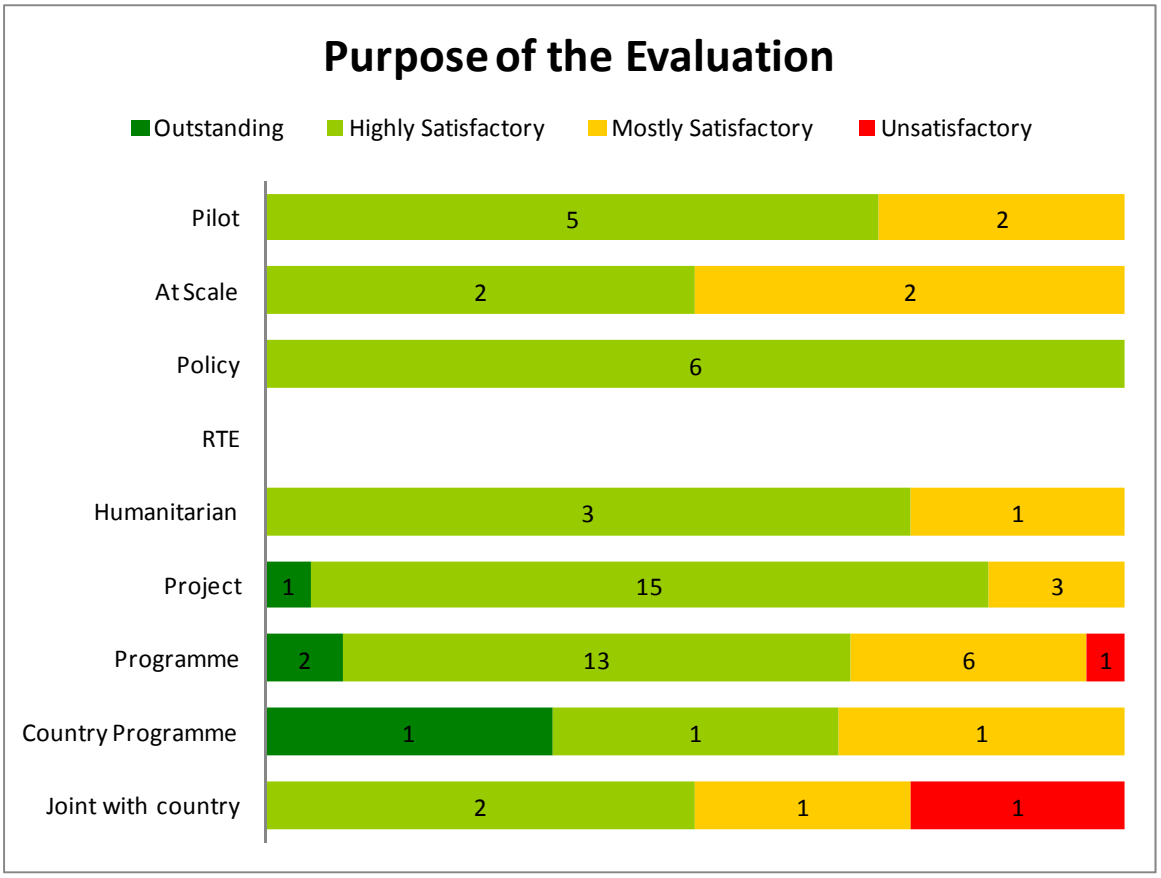

ROSA
Outstanding
Highly Satisfactory
Mostly Satisfactory
Unsatisfactory
UNICEF
Joint with UN
Joint with other
Joint with country
Externally Managed
Not Clear
Country-led
1
10
1
1

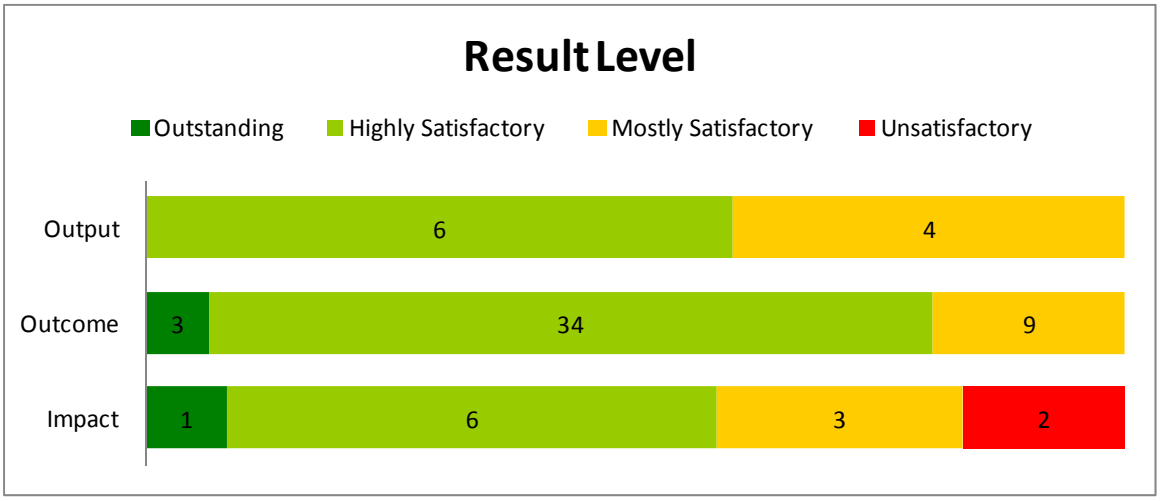

WCARO
Outstanding
Highly Satisfactory
Mostly Satisfactory
Unsatisfactory
UNICEF
Joint with UN
Joint with other
Joint with country
Externally Managed
Not Clear
Country-led
3
4

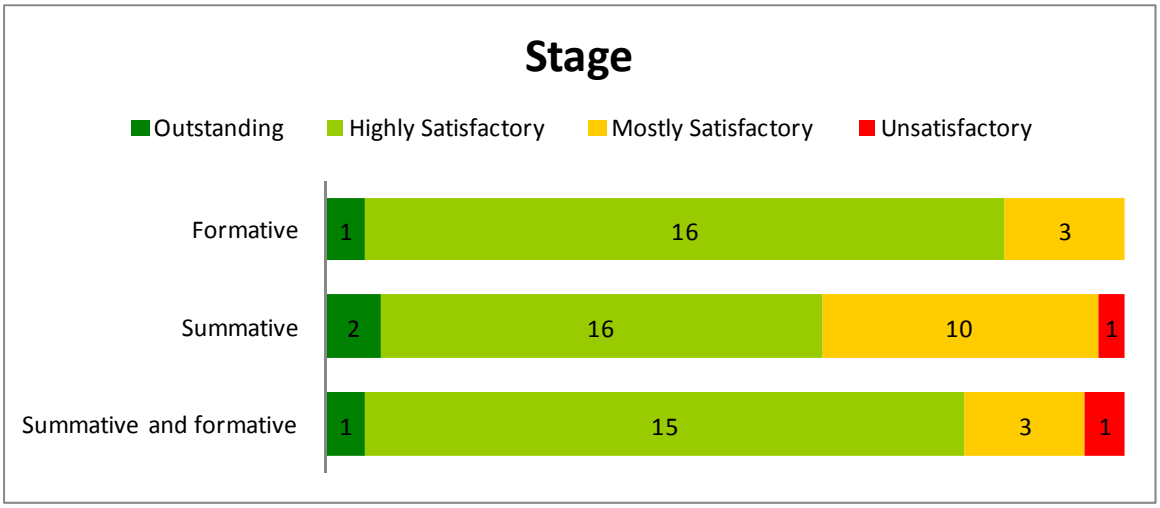

HQ
Outstanding
Highly Satisfactory
Mostly Satisfactory
Unsatisfactory
UNICEF
Joint with UN
Joint with other
Joint with country
Externally Managed
Not Clear
Country-led
9

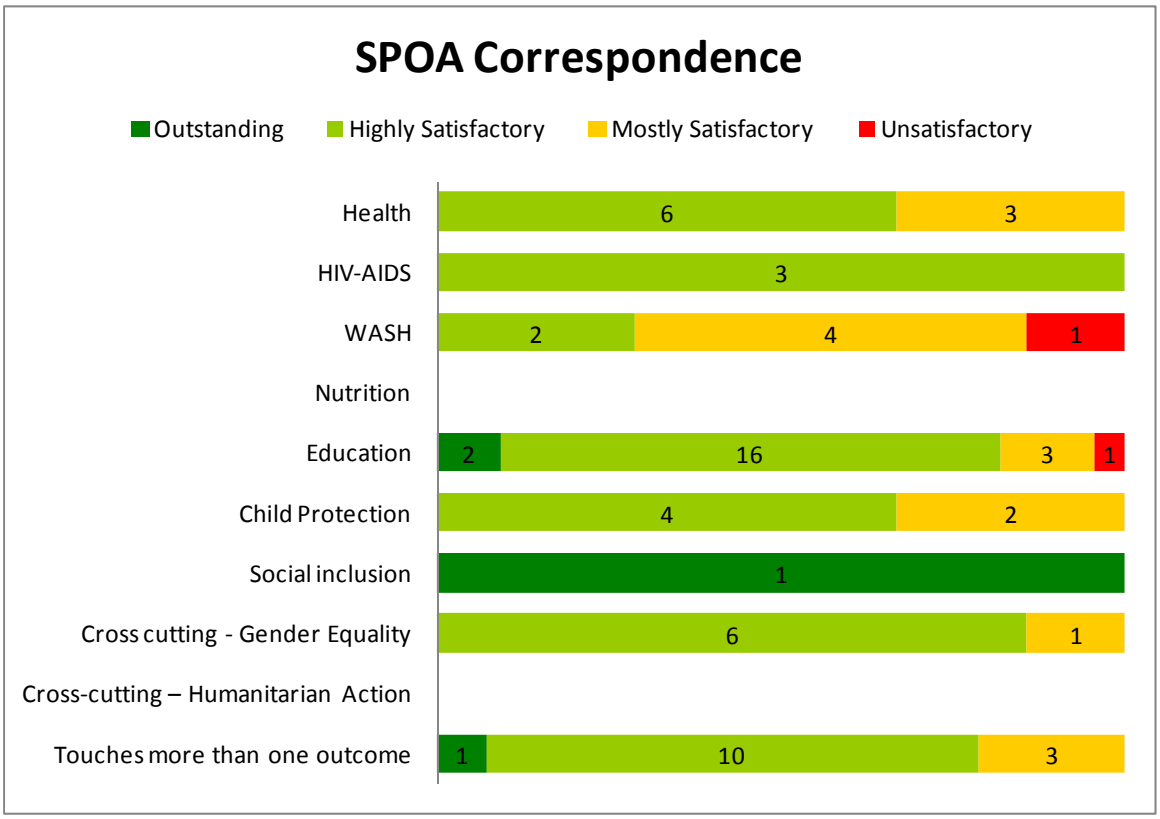
Appendix IX Overall Ratings Graphs by Report Typology

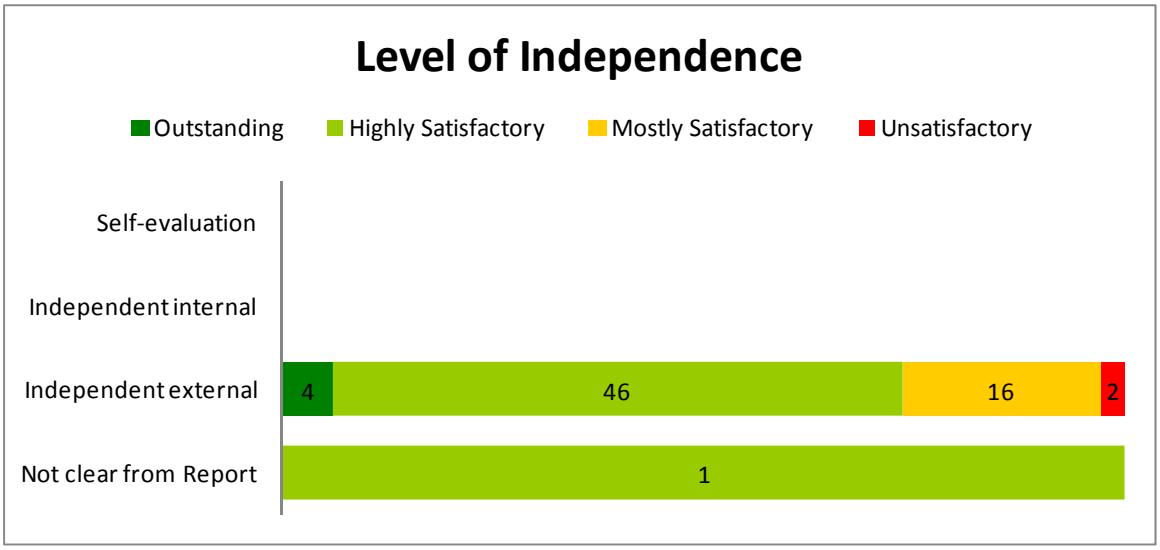


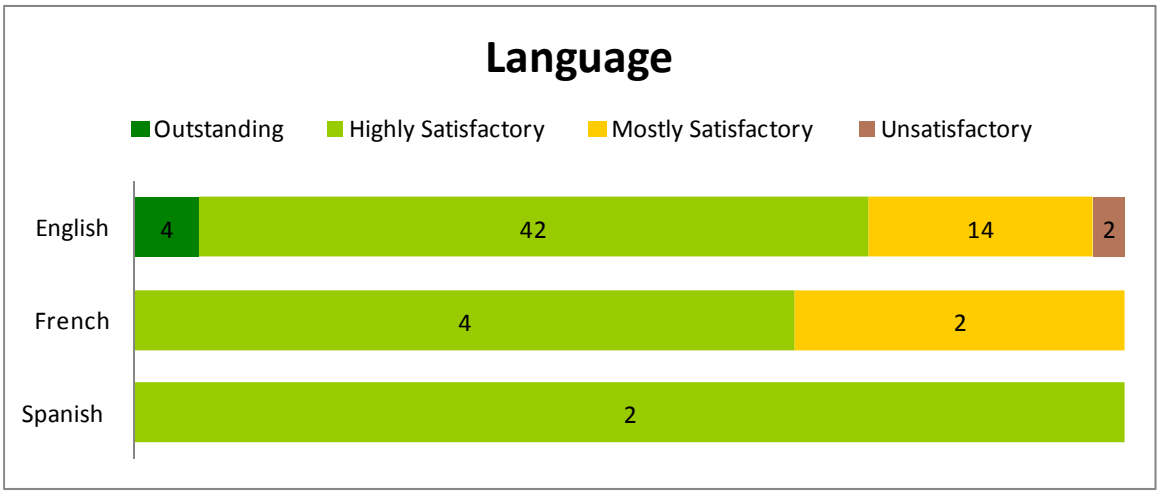

Purpose of the Evaluation
Outstanding
Highly Satisfactory
Mostly Satisfactory
Unsatisfactory
Pilot
5
2
At Scale
2
2
Policy
6
RTE
Humanitarian
3
1
Project
1
15
3
Programme
2
13
6
1
Country Programme
1
1
1
Joint with country
2
1
1


Result Level
Outstanding
Highly Satisfactory
Mostly Satisfactory
Unsatisfactory
Output
6
4
Outcome
3
34
9
Impact
1
6
3
2


Stage
Outstanding
Highly Satisfactory
Mostly Satisfactory
Unsatisfactory
Formative
1
16
3
Summative
2
16
10
1
Summative and formative
1
15
3
1

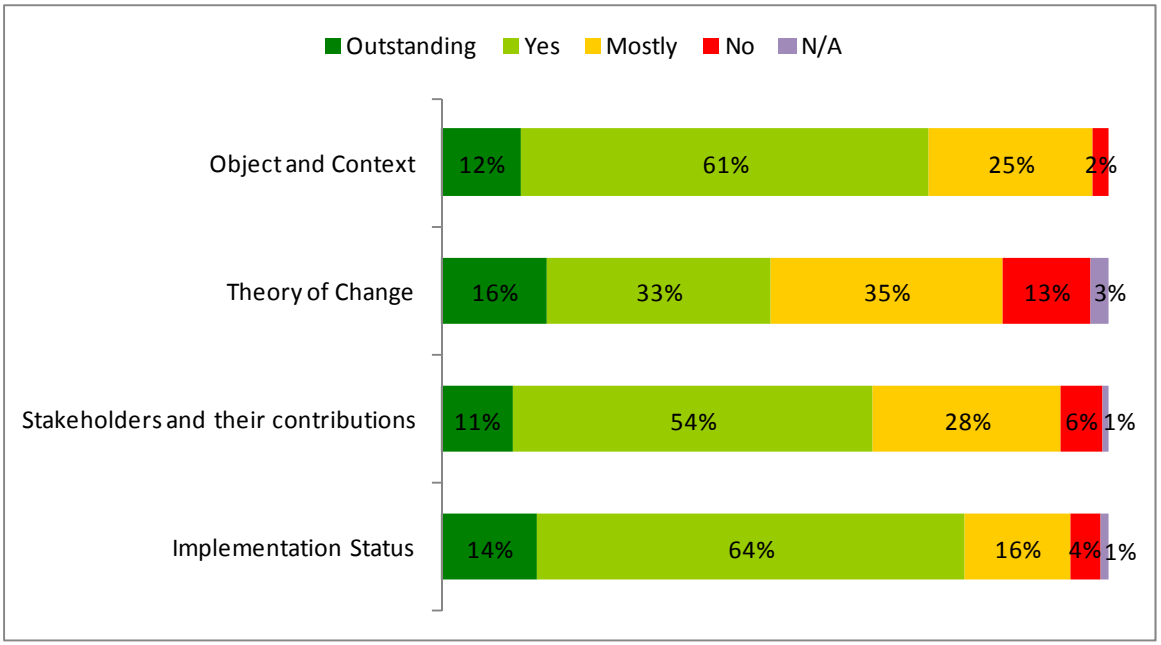

SPOA Correspondence
Outstanding
Highly Satisfactory
Mostly Satisfactory
Unsatisfactory
Health
6
3
HIV-AIDS
3
WASH
2
4
1
Nutrition
Education
2
16
3
1
Child Protection
4
2
Social inclusion
1
Cross cutting - Gender Equality
6
1
Cross-cutting – Humanitarian Action
Touches more than one outcome
1
10
3


Level of Independence
Outstanding
Highly Satisfactory
Mostly Satisfactory
Unsatisfactory
Self-evaluation
Independent internal
Independent external
Not clear from Report
4
46
16
2
1

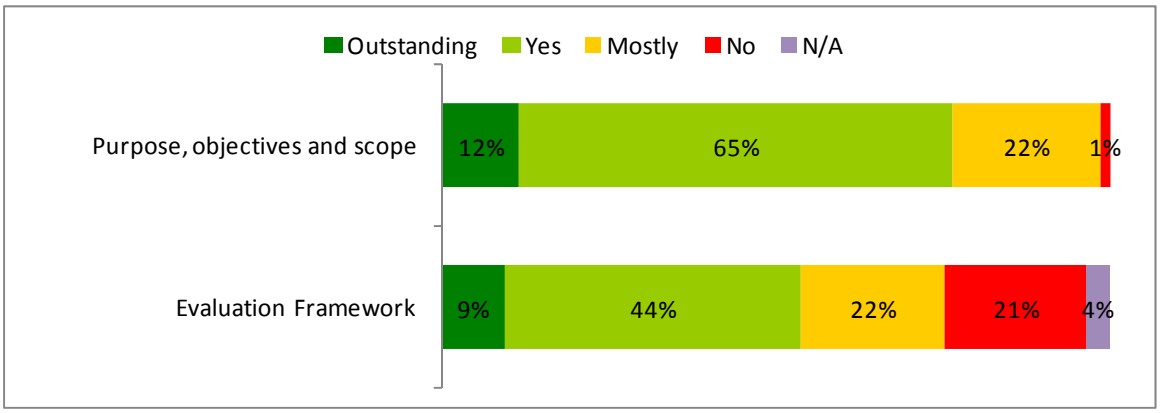

Language
Outstanding
Highly Satisfactory
Mostly Satisfactory
Unsatisfactory
English
French
Spanish
4
42
14
2
4
2
2

Appendix X Sub-Section Ratings

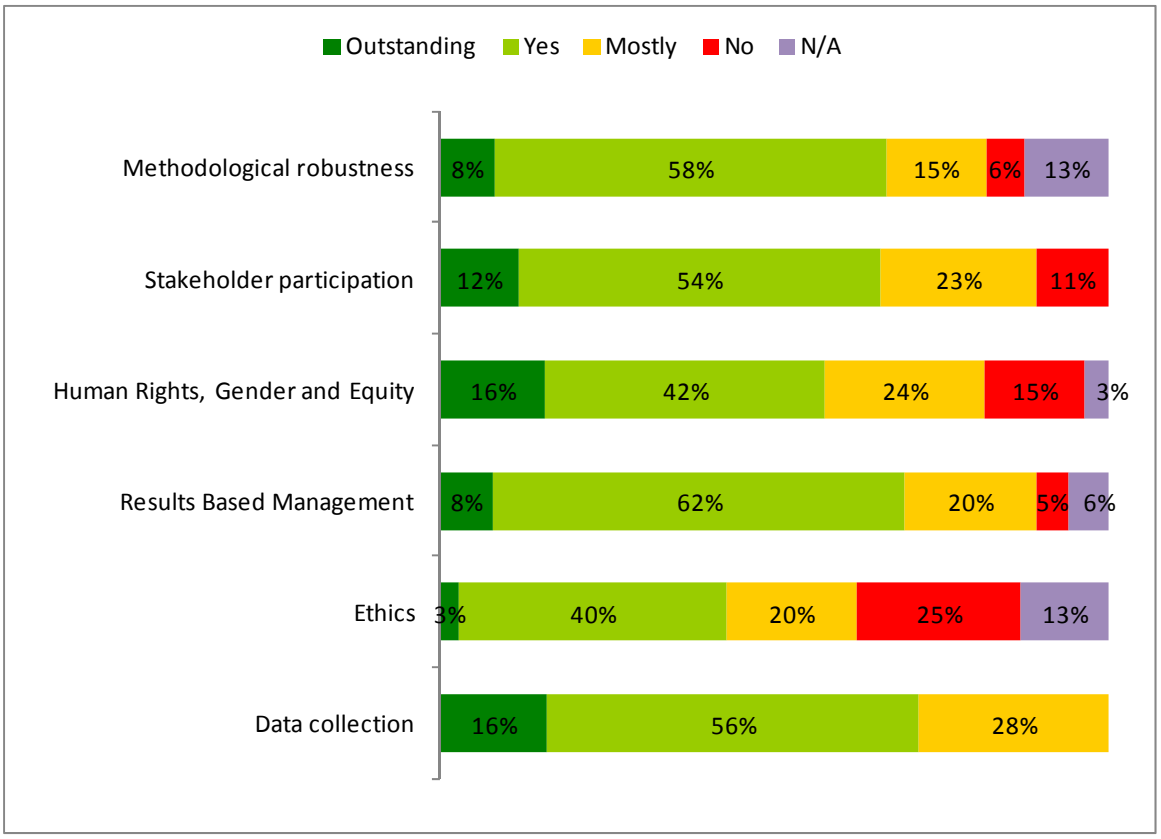
Section A Object of the Evaluation



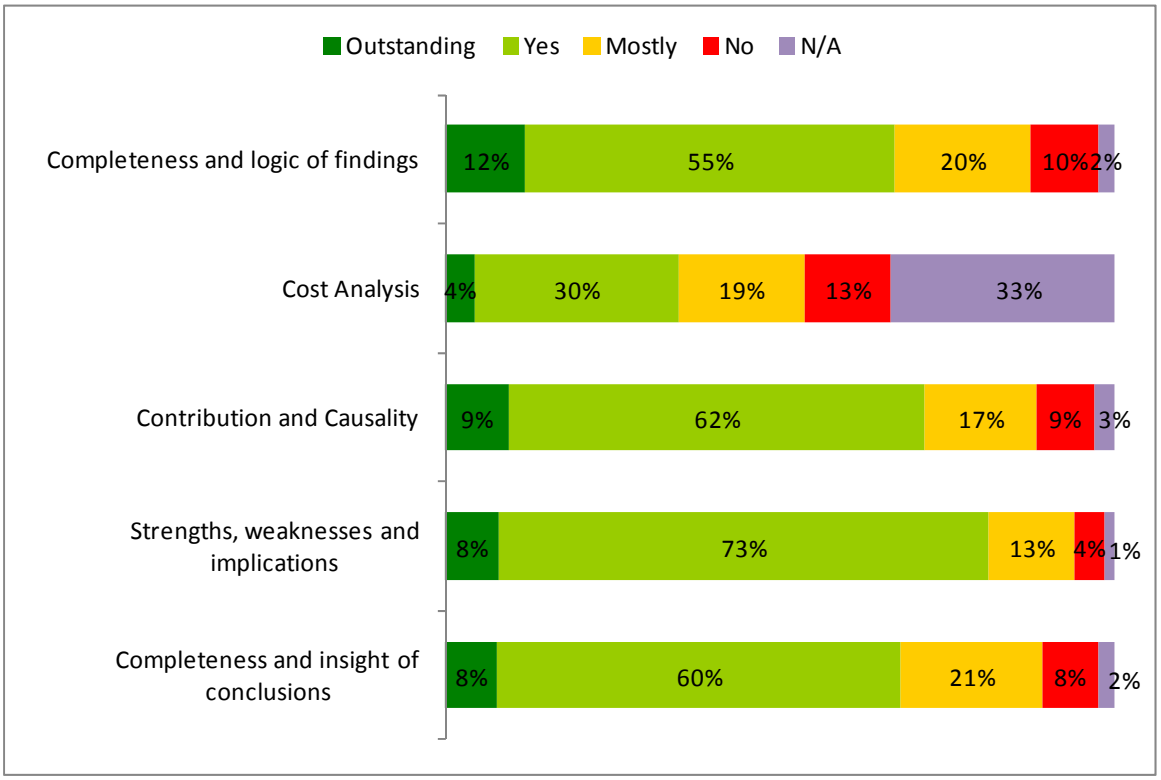
Section B Evaluation Purpose, Objectives and Scope



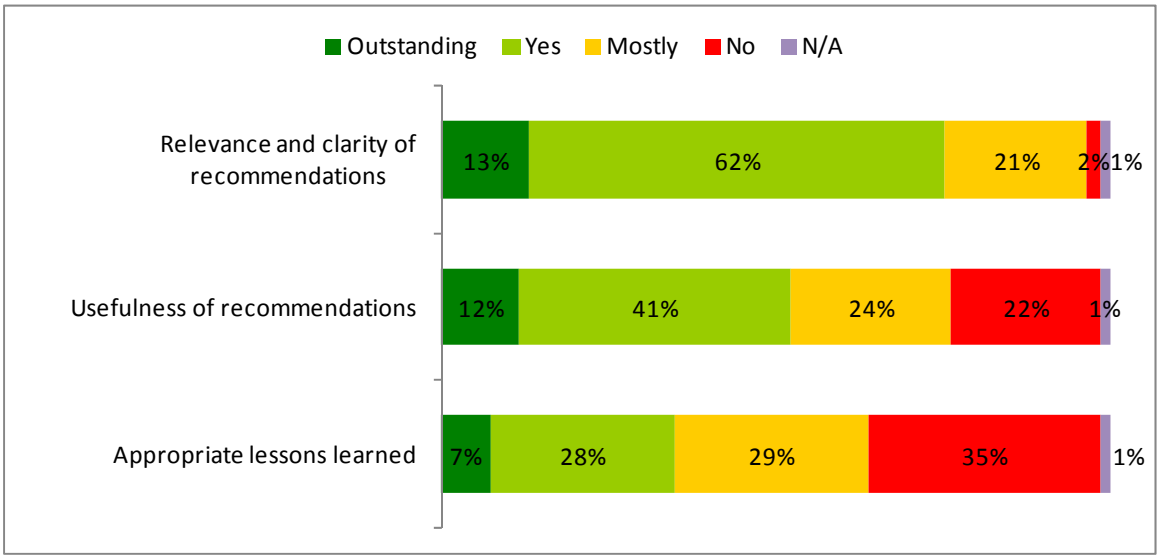
Section C Evaluation Methodology, Gender, Human Rights and Equity



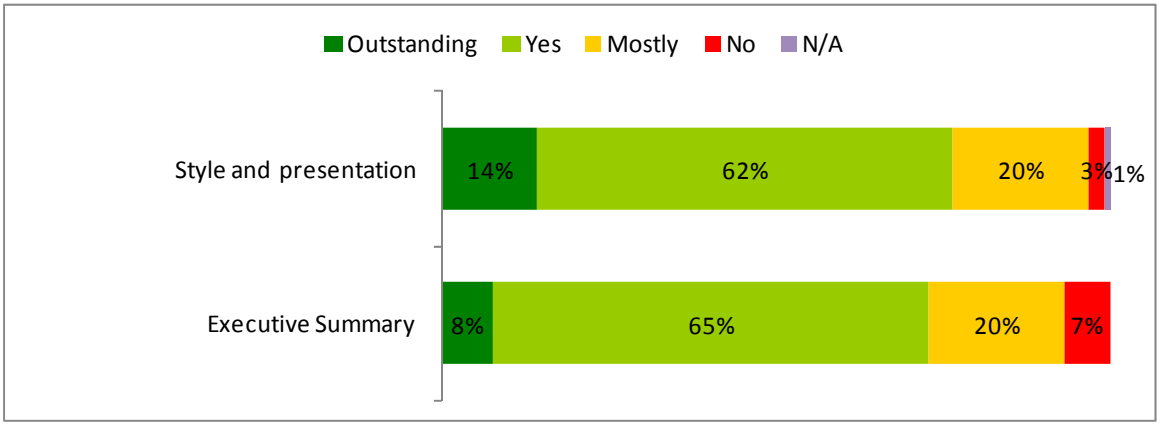
Section D Findings and Conclusions



Section E Recommendations and Lessons Learned



Section F Report is Well Structured, Logic and Clear



Appendix XI Analysis Table for 2014 Reports

Rating Category	Region									Geographic Scope					Management of Evaluation							Purpose									Results			SPOA Correspondence										Level of Independence				Approach		
	Latin America and Caribbean Regional Office	Central & Eastern Europe, Commonwealth of Independent States	East Asia and the Pacific Regional Office	Eastern and Southern Africa Regional Office	Middle East and North Africa Regional Office	South Asia Regional Office	West and Central Africa Regional Office	Corporate (HQ)	Other	1.1 Sub-national	1.2 National	1.3 Multi-country	1.4 Regional	1.5 Multi-region	2.1 UNICEF managed	2.2 Joint managed, with one or more UN agencies	2.3 Joint managed, with organisations outside the UN system	2.4 Jointly Managed with Country	2.5 Country-led Evaluation	2.6 Externally managed	2.7 Not clear from Report	3.1 Pilot	3.2 At scale	3.3 Policy	3.4 Real-time-evaluation	3.5 Humanitarian	3.6 Project	3.7 Programme	3.8 Country Programme Evaluation (CPE)	3.9 Impact Evaluation	4.1 Output	4.2 Outcome	4.3 Impact	5.1 Health	5.2 HIV-AIDS	5.3 WASH	5.4 Nutrition	5.5 Education	5.6 Child Protection	5.7 Social inclusion	5.8 Cross cutting - Gender Equality	5.9 Cross-cutting – Humanitarian Action	5.10 Touches more than one outcome	6.1 Self-evaluation	6.2 Independent internal	6.3 Independent external	6.4 Not clear from Report	7.1 Formative	7.2 Summative	7.3 Summative and formative
OVERALL																																															
Outstanding, best practice	0	3	0	0	0	1	0	0	0	1	2	1	0	0	4	0	0	0	0	0	0	0	0	0	0	0	1	2	1	0	3	1	0	0	0	0	2	0	1	0	0	1	0	0	4	0	1	2	1
Highly Satisfactory	4	10	4	6	1	10	3	9	0	26	15	1	0	5	40	1	1	5	0	0	0	5	2	6	0	3	15	13	1	2	6	34	6	6	3	2	0	16	4	0	6	0	10	0	0	46	1	16	16	15
Mostly Satisfactory	1	1	1	5	2	1	5	0	0	9	6	0	1	0	11	2	1	2	0	0	0	2	2	0	0	1	3	6	1	1	4	9	3	3	0	4	0	3	2	0	1	0	3	0	0	16	0	3	10	3
Unsatisfactory	0	0	0	0	0	1	1	0	0	0	2	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	2	0	0	1	0	1	0	0	0	0	0	0	0	2	0	0	1	1
SECTION A																																															
Outstanding, best practice	2	3	0	1	0	0	0	0	0	1	4	0	1	0	6	0	0	0	0	0	0	2	0	0	0	0	3	0	1	0	2	4	0	0	0	0	0	2	1	1	0	0	2	0	0	6	0	3	1	2
Highly Satisfactory	2	9	3	3	1	9	3	9	0	20	13	1	0	5	35	1	1	2	0	0	0	3	2	6	0	2	10	13	2	1	5	28	5	3	3	2	0	15	4	0	4	0	8	0	0	39	0	12	15	12
Mostly Satisfactory	1	2	2	6	2	4	6	0	0	14	8	1	0	0	14	1	2	6	0	0	0	2	2	0	0	1	6	9	0	3	3	13	7	6	0	4	0	5	1	0	3	0	4	0	0	22	1	5	13	5
Unsatisfactory	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1
SECTION B																																															
Outstanding, best practice	0	3	1	0	0	1	0	0	0	2	3	0	0	0	5	0	0	0	0	0	0	0	0	0	0	1	2	1	1	0	1	3	1	0	0	0	0	2	0	1	1	0	1	0	0	5	0	1	1	3
Highly Satisfactory	3	10	1	6	1	10	2	6	0	21	11	1	1	4	33	1	1	4	0	0	0	5	3	2	0	1	13	13	1	1	5	30	3	8	2	1	0	11	4	0	5	0	8	0	0	38	1	15	16	8
Mostly Satisfactory	2	1	3	5	2	2	7	3	0	13	11	1	0	1	17	2	2	4	0	0	0	2	1	4	0	2	4	8	1	3	4	13	8	1	1	6	0	9	2	0	1	0	5	0	0	25	0	4	12	9
Unsatisfactory	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SECTION C																																															
Outstanding, best practice	0	4	0	0	0	1	0	0	0	1	3	1	0	0	5	0	0	0	0	0	0	0	0	0	0	0	1	3	1	0	0	3	2	0	0	0	0	3	0	1	0	0	1	0	0	5	0	1	2	2
Highly Satisfactory	5	10	3	6	3	8	3	6	0	25	13	1	1	4	37	1	1	5	0	0	0	6	1	4	0	2	14	13	2	2	6	33	5	6	3	1	0	14	5	0	4	0	11	0	0	43	1	16	17	11
Mostly Satisfactory	0	0	2	5	0	3	5	3	0	10	7	0	0	1	13	2	1	2	0	0	0	1	3	2	0	2	4	5	0	1	4	10	3	3	0	5	0	4	1	0	3	0	2	0	0	18	0	3	9	6
Unsatisfactory	0	0	0	0	0	1	1	0	0	0	2	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	2	0	0	1	0	1	0	0	0	0	0	0	0	2	0	0	1	1
SECTION D																																															
Outstanding, best practice	0	1	1	0	0	0	0	3	0	1	0	1	0	3	5	0	0	0	0	0	0	0	0	0	0	0	2	0	3	0	0	2	2	1	0	0	0	1	0	0	2	0	2	0	0	5	0	3	1	1
Highly Satisfactory	4	10	3	8	1	10	4	6	0	25	18	1	0	2	39	1	1	5	0	0	0	3	3	6	0	1	16	13	2	2	3	35	7	7	2	3	0	17	4	1	2	0	10	0	0	45	1	12	19	15

Rating Category	Region									Geographic Scope					Management of Evaluation							Purpose										Results			SPOA Correspondence										Level of Independence				Approach		
	Latin America and Caribbean Regional Office	Central & Eastern Europe, Commonwealth of Independent States Regional Office	East Asia and the Pacific Regional Office	Eastern and Southern Africa Regional Office	Middle East and North Africa Regional Office	South Asia Regional Office	West and Central Africa Regional Office	Corporate (HQ)	Other	1.1 Sub-national	1.2 National	1.3 Multi-country	1.4 Regional	1.5 Multi-region	2.1 UNICEF managed	2.2 Joint managed, with one or more UN agencies	2.3 Joint managed, with organisations outside the UN system	2.4 Jointly Managed with Country	2.5 Country-led Evaluation	2.6 Externally managed	2.7 Not clear from Report	3.1 Pilot	3.2 At scale	3.3 Policy	3.4 Real-time-evaluation	3.5 Humanitarian	3.6 Project	3.7 Programme	3.8 Country Programme Evaluation (CPE)	3.9 Impact Evaluation	4.1 Output	4.2 Outcome	4.3 Impact	5.1 Health	5.2 HIV-AIDS	5.3 WASH	5.4 Nutrition	5.5 Education	5.6 Child Protection	5.7 Social inclusion	5.8 Cross cutting - Gender Equality	5.9 Cross-cutting – Humanitarian Action	5.10 Touches more than one outcome	6.1 Self-evaluation	6.2 Independent internal	6.3 Independent external	6.4 Not clear from Report	7.1 Formative	7.2 Summative	7.3 Summative and formative
Mostly Satisfactory	1	3	1	3	2	2	4	0	0	10	5	0	1	0	11	2	1	2	0	0	0	4	1	0	0	1	3	5	1	1	5	9	2	2	1	3	0	3	2	0	3	0	2	0	0	16	0	5	8	3
Unsatisfactory	0	0	0	0	0	1	1	0	0	0	2	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	2	0	0	1	0	1	0	0	0	0	0	0	0	2	0	0	1	1
SECTION E																																																		
Outstanding, best practice	0	0	0	0	0	1	0	1	0	1	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	1	0	0	1	0	0	0	0	2	0	1	0	1
Highly Satisfactory	2	12	3	4	2	6	1	4	0	15	14	2	0	3	29	2	1	2	0	0	0	1	1	3	0	3	14	10	2	0	2	28	4	3	0	2	0	11	3	1	4	0	10	0	0	33	1	13	10	11
Mostly Satisfactory	3	2	2	7	1	6	7	4	0	19	11	0	1	1	23	1	2	6	0	0	0	6	3	3	0	0	5	10	1	4	7	17	7	6	3	4	0	10	3	0	2	0	4	0	0	32	0	6	18	8
Unsatisfactory	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0
SECTION F																																																	
Outstanding, best practice	0	3	1	0	0	2	0	0	0	2	3	1	0	0	6	0	0	0	0	0	0	0	0	0	0	1	1	2	2	0	1	4	1	0	0	0	0	2	0	1	1	0	2	0	0	6	0	2	2	2
Highly Satisfactory	3	10	3	6	1	10	3	9	0	25	14	0	1	5	37	2	1	5	0	0	0	5	2	6	0	2	14	14	0	2	7	32	5	4	3	2	0	16	5	0	6	0	9	0	0	44	1	15	18	12
Mostly Satisfactory	2	1	1	5	2	1	6	0	0	9	8	1	0	0	12	1	2	3	0	0	0	2	2	0	0	1	4	6	1	2	2	10	6	5	0	5	0	4	1	0	0	0	3	0	0	18	0	3	9	6
Unsatisfactory	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Appendix XII Clarification of Criteria for Completing GEROS Reviews

Report sequence number											
Title of the Evaluation Report											
Year of the Evaluation Report	2016	2015	2014	2013	2012	2011	2010				
Country(ies)											
Region	Latin America and Caribbean Regional Office	Central & Eastern Europe, Commonwealth of Independent States RO	East Asia and the Pacific Regional Office	Eastern and Southern Africa Regional Office	Middle East and North Africa Regional Office	South Asia Regional Office	West and Central Africa Regional Office	Corporate (HQ)	Other		
Date of review											
Type of Report	Evaluation	Needs assessment	Appraisal	Evaluability	Review, including mid-term review	Inspection	Investigation	Research & study	Audit	Survey	Internal Management consulting
Geographic Scope ***Coverage of the programme being evaluated and generalizability of evaluation findings***	1.1 Sub-national: The programme and evaluation covers selected sub-national units (districts, provinces, states, etc.) within a country, where results cannot be generalized to the whole country	1.2 National: The programme covers the whole country, and the evaluation draws a sample in every district, or uses a sampling frame that is representative of the whole country.	1.3 Multi-country: Where one programme is implemented in several countries, or different programmes of a similar theme are implemented in several countries, the evaluation would cover two or more countries	1.4 Regional: Where one programme is implemented in several countries, or different programmes of a similar theme are implemented in several countries, the evaluation covers multiple countries within the region and the sampling is	1.5 Multi-region/Global: The programme is implemented in two or more regions, or deliberately targets all regions. The evaluation would typically sample several countries across multiple						

			within one region. The results of the evaluation would not be generalizable to other countries in the region.	adequate to make the results generalizable to the region.	regions, with the results intended to be generalizable in two or more regions.						
Management ***Managerial control and oversight of evaluation decisions (i.e., TORs, selection of consultants, budgets, quality assurance and approval of evaluation findings).In all instances, it is assumed that the management approaches include relevant national actors (e.g., government, universities, NGOs, CBOs)***	2.1 UNICEF managed: Working with national partners of different categories UNICEF is responsible for all aspects of the evaluation.	2.2 Joint managed, with one or more UN agencies: UNICEF is the co-manager with one or more UN agencies	2.3 Joint managed, with organisations outside the UN system: UNICEF is the co-manager with one or more organizations outside the UN system	2.4. Jointly Managed with Country: Evaluations jointly managed by the Country (Government and/or CSO) and the UNICEF CO	2.5. Country-led Evaluation: Evaluations managed by the Country (Government and/or CSO)	2.6 Externally managed: An external organization manages the evaluation, where UNICEF is one of the organizations being assessed (UN and non-UN)	2.7 Not clear from Report				
Purpose ***Speaks to the overarching goal for conducting the evaluation; its raison d'etre***	3.1 Pilot: Where a new solution, approach, or programme is being tested at a national or sub-national level, the evaluation examines the efficacy of such an intervention with the intention to determine suitability for scaling-up.	3.2 At scale: The evaluation examines the efficacy of a programme that is being implemented at or near its maximum intended extent, with the intention of providing feedback on efficiency and the overall effectiveness of the programme to scale up focus for lessons	3.3 Policy: An evaluation whose main purpose is to examine the results of a policy that is delinked from field-based programming operations.	3.4 Real-time-evaluation: In the context of an emergency, an evaluation of the efficacy of the response, which collates lessons that can be applied back to an on-going response	3.5 Humanitarian: Humanitarian evaluation assesses organizational performance in emergency settings (including both natural disasters & conflicts) at various phases of these crises, from preparedness and risk	3.6 Project: An evaluation which is step-by-step process of collecting, recording and organisation information about the project results including immediate results, short-term outputs and long-term project	3.7 Programme: An evaluation of a sectorial programme to determine its overall effectiveness and efficiency in relation to the stated goals and objectives	3.8 Country Programme Evaluation (CPE): An evaluation that assess the relevance, effectiveness, efficiency, sustainability of the entire UNICEF Country Programme	3.9 Impact Evaluation: An evaluation that looks at the positive and negative, primary and secondary long-term effects on final beneficiaries produced by a development intervention. Impact evaluations assess the direct and indirect contributions of the		

		learned.			reduction to response, recovery & the transition to development	outcomes			intervention to specific development results, using robust quantitative, qualitative, or mixed methods to assign contribution to higher level results.		
Result ***Level of changes sought, as defined in results based management: refer to substantial use of highest level reached***	4.1 Output: Causal effects deriving directly from programme activities, and assumed to be completely under programme control	4.2 Outcome: Effects from one or more programmes being implemented by multiple actors (UNICEF and others), where the cumulative effect of outputs elicits results beyond the control of any one agency or programme	4.3 Impact: Final results of a programme or policy on the intended beneficiaries and, where possible, on comparison groups. Reflects the cumulative effect of donor supported programmes of cooperation and national policy initiatives.								
SPOA Correspondence ***Alignment with SPOA focus area priorities: (1) Young child survival and development; (2) Basic education and gender equality; (3) HIV/AIDS and children; (4) Child protection from violence, exploitation and abuse; and (5) Policy***	5.1 Health: Supporting global efforts to reduce under-five mortality through improved and equitable use of high impact maternal, newborn and child health interventions from pregnancy to adolescence and promotion of healthy behaviours. Programme areas: a) Immunization b) Polio eradication c) Maternal and newborn health d) Child health e) Health systems strengthening f) Health in	5.2 HIV-AIDS: Supporting global efforts to prevent new HIV infections and increase treatment during both decades of a child's life through improved and equitable use of proven HIV prevention and treatment interventions by pregnant women, children and	5.3 WASH: Supporting global efforts to eliminate open defecation and increase use of safe drinking water through improved and equitable access to safe drinking water sources, sanitation and healthy environment	5.4 Nutrition: Supporting global efforts to reduce undernutrition, with particular focus on stunting, through improved and equitable use of nutritional support and improved nutrition and care practices. Programme areas: a) Infant and young child	5.5 Education: Supporting global efforts to provide access to quality education for both boys and girls through improved learning outcomes and equitable and inclusive education. Programme areas: a) Early learning	5.6 Child Protection: Supporting global efforts to prevent violence, abuse, exploitation and neglect through improved and equitable prevention and child protection systems. Programme areas: a) Child protection	5.7 Social inclusion: Supporting global efforts to reduce child poverty and discrimination against children through improved policy environments and systems for disadvantaged children. Programme areas: a) Child poverty	5.8 Cross cutting - Gender Equality: UNICEF will emphasize the empowerment of girls and women and address gender-related needs of girls, boys, fathers, mothers and communities. UNICEF will identify and leverage positive	5.9 Cross-cutting – Humanitarian Action: UNICEF will strive to save lives and protect rights as defined in the Core Commitments for Children in Humanitarian Action in line with internationally accepted standards. UNICEF will focus its efforts on	5.10 Touches more than one outcome. Please specify in the comments.	

advocacy and partnerships for children's rights	humanitarian situations	adolescents. Programme areas: a) Prevention of mother to child transmission and infant male circumcision b) Care and treatment of young children affected by HIV&AIDS c) Adolescents and HIV&AIDS d) Protection and support for children and families e) HIV-AIDS in humanitarian situations	s and improved hygiene practices. Programme areas: a) Water supply b) Sanitation c) Hygiene d) WASH in schools and early childhood developments centers e) WASH in humanitarian situations	feeding b) Micronutrients c) Nutrition and HIV d) Community management of acute malnutrition e) Nutrition in humanitarian situations	b) Equity with a focus on girls' education and inclusive education c) Learning and child friendly schools d) Education in humanitarian situations	systems strengthening b) Violence, exploitation and abuse c) Justice for children d) Birth registration e) Strengthened families and communities f) Child protection in humanitarian situations	analysis and social protection b) Human rights, non-discrimination and participation c) Public finance management d) Governance and decentralization e) Social inclusion in humanitarian situations	cross-sectoral synergies and linkages such as those among improving girls' education, ending child marriage and reducing maternal mortality. UNICEF will also focus on increasing access to services and opportunities by women and girls and their inclusion and participation in all facets of life as well as on advocacy and technical support on gender-equitable policies, budgeting and resource allocations. Emphasis will be placed on: a) Sex-disaggregated and other gender related data. b) Promote gender-sensitive interventions as a core programmatic priority, c) To the extent possible, all relevant policies, programmes	systematically reducing vulnerability to disasters and conflicts for effective prevention of and response to humanitarian crises, on improving links between development programmes and humanitarian response and on promoting rapid recovery and building community resilience to shocks that affect children. Emphasis on: a) Support for humanitarian action – achieving faster scaling up of response b) Early identification of priorities and strategies c) Rapid deployment of qualified staff and clear accountabilities d) Responses consistent with humanitarian principles in situations of unrest or armed conflict		

								and activities will mainstream gender equality.			
Level of Independence ***Implementation and control of the evaluation activities***	6.1 Self-evaluation: A significant component of evaluation management activities & decision-making about the evaluation are implemented by individuals associated with the target programme/intervention (eg. programmes officer/specialists)	6.2 Independent internal: The evaluation is implemented by consultants but managed in-house by UNICEF professionals. The overall responsibility for the evaluation lies within the division whose work is being evaluated.	6.3 Independent external: The evaluation is implemented by external consultants and/or UNICEF Evaluation Office professionals. The overall responsibility for the evaluation lies outside the division whose work is being evaluated.	6.4 Not clear from Report							
Approach	7.1 Formative: An evaluation with the purpose and aim of improving the programme. Formative evaluations strengthen or improve the object being evaluated by examining the delivery of the programme	7.2 Summative: An evaluation that examines the effects or outcomes of the object being evaluated and summarize it by describing what happened subsequent to delivery of the programme	7.3 Summative and formative: An evaluation that combines the elements of a formative and a summative evaluation.								
TORs present	Yes	No									
Question Criteria	Outstanding	Yes	Mostly	No	N/A						
Section Rating Criteria	Outstanding, best practice	Highly satisfactory	Mostly Satisfactory	Unsatisfactory							

Appendix XIII Special Cases in the 2014 GEROS Exercise

The 2014 GEROS exercise included a number of reports that were slightly different from traditional evaluations. Some challenges arose when it came time to assess these reports according to the GEROS criteria, which raised some questions regarding the flexibility of the template. Below is a list of the special cases found among the 2014 report population, as well as the reasons why the template was sometimes difficult to apply.

Case Studies and Global Syntheses

GEROS-Afghanistan-2014-001	GEROS-Bangladesh-2014-002
GEROS-Brazil-2014-001	GEROS-Cambodia-2014-001
GEROS-Zimbabwe-2014-002	

In all, 5 case studies were included among the reports reviewed in 2013. These reports were part of larger, global synthesis evaluations. As requested by UNICEF, each report (the synthesis report and all case studies) was assessed individually according to the GEROS criteria. However, it was unclear if evaluators were aware that their case studies would be analysed as separate reports (i.e. so they could take appropriate measures to include the necessary detail in each one, prior to submission).

The case studies analysed this year all received highly satisfactory ratings, notably because they were detailed and quite comprehensive. They included information on the context, methodology and analysis. Nevertheless, case studies do not always include such depth, nor do they necessarily require a description of the evaluation framework or a set of conclusions and recommendations (normally provided in the global synthesis report). On the other hand, global syntheses may not always contain specific details or examples from the countries evaluated. UNICEF may wish to consider how to deal with such reports in future, and whether a case study can realistically be judged against the same criteria as a full evaluation report.

Impact Evaluations

	Rating
GEROS-Bangladesh-2014-014	Highly satisfactory
GEROS-Lesotho-2014-001	Highly satisfactory
GEROS-Nigeria-2014-012	Mostly satisfactory
GEROS-Sierra_Leone-2014-003	Unsatisfactory

Four impact evaluations were included among the GEROS reports this year. These reports received a mixed set of ratings, from highly satisfactory to unsatisfactory.

According to reviewers, the template was difficult to apply to these impact evaluations, notably because of the way evaluation criteria were used. It is generally accepted that impact evaluations are not required to address all of the OECD/DAC criteria (only impact), so this choice need not be justified within the description of the evaluation. Further, due to their narrow focus, the impact evaluations reviewed provided little information around results-based management and the evaluated object's theory of change. Other issues related to how recommendations and lessons learned should be judged, given the unique nature of impact evaluations. In many cases, therefore, GEROS template criteria simply appeared inapplicable.

Future Considerations for UNICEF

Given these special cases in the 2014 review sample, UNICEF may wish to consider modifying its template or finding new ways of assessing unique reports. For instance, UNICEF could:

- Adapt its current template to account for possible variation in content, structure and focus within reports. Naturally, caution should be taken to prevent the template from becoming so flexible that common standards are easy to circumvent;
- Create a new template for each different type of report. The comparability of all the reports (for meta-analysis purposes) would then likely require further thought or discussion, especially if certain questions were removed or altered in some versions of the template. In this case, UNICEF may wish to conduct separate analyses of each report type, comparing them to each other rather than to the entire population of reports;
- Leave the template as-is, even if future good quality reports may receive a lower rating due to the absence/special treatment of some criteria.