An Independent Review of
UNICEF Evaluation Report Quality
and Selected Trends from 2016 – 2018

FINAL REPORT

May 2018 – v. 4
Principal Author: Katherine Garven
Contributing Authors: Fernando Garabito and Esther Rouleau

Evaluation Managers: Uyen Kim Huynh and Lovemore Mhuriyengwe
## Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>COs</td>
<td>Country Offices</td>
</tr>
<tr>
<td>EAPR</td>
<td>East Asia and Pacific Region</td>
</tr>
<tr>
<td>ECAR</td>
<td>Europe and Central Asia Region</td>
</tr>
<tr>
<td>EO</td>
<td>Evaluation Office</td>
</tr>
<tr>
<td>ESAR</td>
<td>Eastern and Southern Africa Region</td>
</tr>
<tr>
<td>GEROS</td>
<td>Global Evaluation Report Oversight System</td>
</tr>
<tr>
<td>GEEW</td>
<td>Gender Equality and Empowerment of Women</td>
</tr>
<tr>
<td>HQ</td>
<td>Headquarters</td>
</tr>
<tr>
<td>HRBAP</td>
<td>Human Rights Based Approach to Programming</td>
</tr>
<tr>
<td>LACR</td>
<td>Latin America and Caribbean Region</td>
</tr>
<tr>
<td>M&amp;E</td>
<td>Monitoring and Evaluation</td>
</tr>
<tr>
<td>MENA</td>
<td>Middle East and North Africa Region</td>
</tr>
<tr>
<td>MTSP</td>
<td>Medium Term Strategic Plan</td>
</tr>
<tr>
<td>N/A</td>
<td>Not Applicable</td>
</tr>
<tr>
<td>OECD/DAC</td>
<td>Organization for Economic Co-operation and Development/Development Assistance Committee</td>
</tr>
<tr>
<td>RBM</td>
<td>Results-based Management</td>
</tr>
<tr>
<td>ROs</td>
<td>Regional Offices</td>
</tr>
<tr>
<td>ROSA</td>
<td>Regional Office of South Asia</td>
</tr>
<tr>
<td>RTE</td>
<td>Real-time evaluation</td>
</tr>
<tr>
<td>SPOA</td>
<td>Strategic Plan Objective Area</td>
</tr>
<tr>
<td>ToC</td>
<td>Theory of Change</td>
</tr>
<tr>
<td>TORs</td>
<td>Terms of Reference</td>
</tr>
<tr>
<td>UN</td>
<td>United Nations</td>
</tr>
<tr>
<td>UNDAF</td>
<td>United Nations Development Assistance Framework</td>
</tr>
<tr>
<td>UNEG</td>
<td>United Nations Evaluation Group</td>
</tr>
<tr>
<td>UNICEF</td>
<td>United Nations Children’s Fund</td>
</tr>
<tr>
<td>UN-SWAP</td>
<td>UN System-wide Action Plan for gender equality and empowerment of women</td>
</tr>
<tr>
<td>WASH</td>
<td>Water, Sanitation and Hygiene</td>
</tr>
<tr>
<td>WCAR</td>
<td>West and Central Africa Region</td>
</tr>
</tbody>
</table>
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acronyms</td>
<td>ii</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>iii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>iv</td>
</tr>
<tr>
<td>List of Annexes</td>
<td>vi</td>
</tr>
<tr>
<td>(under separate cover)</td>
<td>vi</td>
</tr>
<tr>
<td>Executive Summary</td>
<td>vii</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2. Background on the UNICEF GEROS System</td>
<td>1</td>
</tr>
<tr>
<td>3. Purpose, Scope and Objective of GEROS</td>
<td>2</td>
</tr>
<tr>
<td>4. Methodology</td>
<td>2</td>
</tr>
<tr>
<td>5. Findings</td>
<td>4</td>
</tr>
<tr>
<td>5.1 Trends within the UNICEF Evaluation Portfolio</td>
<td>5</td>
</tr>
<tr>
<td>5.2 Overall Evaluation Report Quality from 2016 - 2018</td>
<td>1</td>
</tr>
<tr>
<td>5.3 Evaluation Report Quality per Assessment Criterion from 2016 - 2018</td>
<td>7</td>
</tr>
<tr>
<td><strong>Section A: Background</strong></td>
<td>8</td>
</tr>
<tr>
<td><strong>Section B: Evaluation Purpose, Objectives and Scope</strong></td>
<td>11</td>
</tr>
<tr>
<td><strong>Section C: Evaluation Methods</strong></td>
<td>13</td>
</tr>
<tr>
<td><strong>Section D: Evaluation Findings</strong></td>
<td>16</td>
</tr>
<tr>
<td><strong>Section E: Evaluation Conclusions and Lessons Learned</strong></td>
<td>19</td>
</tr>
<tr>
<td><strong>Section F: Recommendations</strong></td>
<td>21</td>
</tr>
<tr>
<td><strong>Section G: Evaluation Structure and Presentation</strong></td>
<td>24</td>
</tr>
<tr>
<td><strong>Section H: Evaluation Principles</strong></td>
<td>27</td>
</tr>
<tr>
<td><strong>SWAP Assessment</strong></td>
<td>29</td>
</tr>
<tr>
<td><strong>Section I: Executive Summary</strong></td>
<td>33</td>
</tr>
<tr>
<td>6. Conclusions and Recommendations</td>
<td>38</td>
</tr>
</tbody>
</table>
List of Figures

Exhibit 1. Reports included in GEROS from 2011 – 2018
Exhibit 2: Number of GEROS reports per region and HQ from 2011 - 2018
Exhibit 3: SPOA Correspondence from 2016 - 2018
Exhibit 4: Correspondence with UNICEF 2018-2021 Strategic Plan Goals and Cross-Cutting Themes
Exhibit 5: Evaluation objects from 2016 – 2018
Exhibit 6: Evaluation types from 2016 - 2018
Exhibit 7: Trends in evaluation approach from 2016 - 2018
Exhibit 8: Trends in evaluation design from 2016 - 2018
Exhibit 9: Trends in evaluation level from 2016 - 2018
Exhibit 10: Trends in geographic level from 2016 - 2018
Exhibit 11: Trends in geographic level from 2016 - 2018
Exhibit 12: Overall Evaluation Score from 2016 - 2018
Exhibit 13: Percentage of Reports per Assessment Rating from 2016 – 2018
Exhibit 14: Percentage of Reports per Sub-Grouping of Assessment Scores from 2016 – 2018
Exhibit 15: Overall Evaluation Score from 2016 – 2018 by Region
Exhibit 16: Overall Evaluation Score by Rating from 2016 – 2018 by Region
Exhibit 17: Overall Evaluation Score on a Scale from 2016 – 2018 by SPOA Correspondence
Exhibit 18: Overall Evaluation Report Quality per Assessment Criterion from 2016 – 2018
Exhibit 19: Average Evaluation Report Quality for the “Background” Assessment Criterion from 2016 – 2018
Exhibit 20: Average Evaluation Report Quality per Region for the “Background” Assessment Criterion from 2016 – 2018
Exhibit 21: Average Evaluation Report Quality per Question for the “Background” Assessment Criterion from 2016 – 2018
Exhibit 22: Average Evaluation Report Quality for the “Purpose, Objectives and Scope” Assessment Criterion from 2016 – 2018
Exhibit 23: Average Evaluation Report Quality per Region for the “Purpose, Objectives and Scope” Assessment Criterion from 2016 – 2018
Exhibit 24: Average Evaluation Report Quality per Question for the “Purpose, Objectives and Scope” Assessment Criterion from 2016 – 2018
Exhibit 26: Average Evaluation Report Quality per Region for the “Methods” Assessment Criterion from 2016 – 2018

Exhibit 27: Average Evaluation Report Quality per Question for the “Methods” Assessment Criterion from 2016 – 2018

Exhibit 28: Average Evaluation Report Quality for the “Findings” Assessment Criterion from 2016 – 2018

Exhibit 29: Average Evaluation Report Quality per Region for the “Findings” Assessment Criterion from 2016 – 2018

Exhibit 30: Average Evaluation Report Quality per Question for the “Findings” Assessment Criterion from 2016 – 2018

Exhibit 31: Average Evaluation Report Quality for the “Conclusions and Lessons Learned” Assessment Criterion from 2016 – 2018

Exhibit 32: Average Evaluation Report Quality per Region for the “Conclusions and Lessons Learned” Assessment Criterion from 2016 – 2018

Exhibit 33: Average Evaluation Report Quality per Question for the “Conclusions and Lessons Learned” Assessment Criterion from 2016 – 2018

Exhibit 34: Average Evaluation Report Quality for the “Recommendations” Assessment Criterion from 2016 – 2018

Exhibit 35: Average Evaluation Report Quality per Region for the “Recommendations” Assessment Criterion from 2016 – 2018

Exhibit 36: Average Evaluation Report Quality per Question for the “Recommendations” Assessment Criterion from 2016 – 2018

Exhibit 37: Average Evaluation Report Quality for the “Structure and Presentation” Assessment Criterion from 2016 – 2018

Exhibit 38: Average Evaluation Report Quality per Region for the “Structure and Presentation” Assessment Criterion from 2016 – 2018

Exhibit 39: Average Evaluation Report Quality per Question for the “Structure and Presentation” Assessment Criterion from 2016 – 2018


Exhibit 41: Average Evaluation Report Quality per Region for the “Evaluation Principles” Assessment Criterion from 2016 – 2018


Exhibit 43: Trends in Overall SWAP Scores and per Micro-Criterion from 2016 – 2018

Exhibit 44: Trends in SWAP Scores per Region from 2016 – 2018

Exhibit 45: Trends in SWAP Scores per SWAP Criterion by Region from 2016 – 2018

Exhibit 46: Average Evaluation Report Quality for the “Executive Summary” Assessment Criterion from 2016 – 2018

Exhibit 47: Average Evaluation Report Quality per Region for the “Executive Summary” Assessment Criterion from 2016 – 2018
List of Annexes

(under separate cover)

Annex I ToRs
Annex II GEROS Assessment Criteria
Annex III List of 2018 reports with overall assessment scores
Annex IV Qualitative strengths and weaknesses per assessment criterion with examples
Annex V Good practice examples
Executive Summary

Introduction and Background on the UNICEF GEROS System

The 2018 GEROS Meta-Analysis presents trends in UNICEF evaluation quality from 2016 – 2018 based on evaluation ratings and qualitative feedback emerging from the Global Evaluation Reports Oversight System (GEROS). GEROS, which was established in 2010, provides a holistic system to ensure good quality and credibility of evaluation reports across the decentralized organisation; supports the strengthening of the evaluation function to fully meet UNEG standards; and serves as a platform for institutional learning and the use of evaluative evidence.

The UNICEF Evaluation Office commissioned this report to inform key audiences, which include: The Global Evaluation Committee, UNICEF senior management (at the global, regional and country levels), the UNICEF Evaluation Office, regional evaluation officers, and evaluation managers. Its purpose is to monitor progress, analyse strengths, and identify current challenges and areas for improvement in order to contribute towards the continued strengthening of the UNICEF evaluation function and UNICEF programming.

Methodology

The GEROS process includes a review by an external independent team of evaluation experts of the quality of UNICEF evaluation reports (at the country, regional and global levels) and a yearly meta-analysis. UNICEF-adapted UNEG standards are used as the criteria of quality assessment and inter-related reliability is managed through a quality control process.

This meta-analysis was conducted from May-June 2019. Quantitative data was compiled regarding scores for different aspects of the reports from 2016 - 2018 using Excel. The analysis was carried out across multiple axes, including descriptive elements of the reports; overall trends in quality (across time, assessment criteria (including UN-SWAP performance), and regions. A qualitative assessment was also provided for 2018 trends by quality assessment criteria to identify strengths and areas for improvement. The report also includes a section on GEROS as a Learning Tool that presents insights from the external independent evaluation team around the GEROS tool and how it can be further strengthened to better support the evaluation function and organisational learning. This analysis should be read along with the findings and conclusions of the 2019 GEROS Review that provides a more in-depth assessment of the GEROS system, including how evaluation report quality is defined and assessed.

Summary of Key Findings

1 Evaluation quality is defined according to the GEROS rating assigned to each evaluation report. The GEROS rating is the result of a quality assessment based on OECD/DAC evaluation standards. Reports rated as “satisfactory or highly satisfactory” can be used with confidence to inform decision-making; those rated as fair can be used with caution as substantial improvements could be made to the report; and reports rated as “unsatisfactory” do not meet OECD/DAC evaluation standards and can therefore not be used for decision-making.

2 The GEROS review process assigns an overall rating to each evaluation report on a scale from 0 – 4 and uses a ranking system to classify the quality of reports as follows: unsatisfactory (0.0 – 1.49); fair (1.5 – 2.49); satisfactory (2.5 – 3.49); and highly satisfactory (3.5 – 4.0). The fair and satisfactory categories are further broken down into lower fair (1.5 – 1.99); upper fair (2.0 – 2.49); lower satisfactory (2.5 – 2.99); and upper satisfactory (3.0 – 3.49). For this GEROS Meta-Analysis, changes in percentage score are in relation to the maximum potential score of 4/4 (100 percent).
Trends within the UNICEF Evaluation Portfolio

The breakdown of thematic areas within the UNICEF Portfolio has stayed relatively stable across the three-year period with the health sector comprising the largest number of evaluations in both 2017 and 2018. Across 2016 – 2018, programmes were the most common objects evaluated, with the percentage of programme evaluations increasing to nearly half of the overall evaluation portfolio in 2018. Strategy, pilot / innovation, and system evaluations were the most common irregular objects evaluated in 2018.

Across 2016 – 2018, approximately three fourths of all evaluations were focused on learning (either formative or formative and summative) and drew on mixed quantitative and qualitative methods. The percentage of reports relying exclusively on qualitative methods decreased by 13 percent from 2016 - 2018, indicating an increasing focus on including quantitative methods in evaluation approaches.

The most common evaluation designs used during the three-year period were quasi-experimental, theory-based, participatory, and case study designs. Over the three-year period, over 94 percent of evaluations were targeted at the outcome or impact levels, with nearly half of all evaluations including impact level results.

Over 75 percent of evaluations were commissioned at the country level and were national in scope across 2016 - 2018. Over three fourths of all evaluations were managed exclusively by UNICEF while the percentage of evaluations jointly managed with other UN organisation remained static across 2016 - 2018 at only 2 percent.

Overall Evaluation Report Quality from 2016 – 2018

As Exhibit A indicates, the average quality of UNICEF evaluation reports was considered as satisfactory across the three-year period and improved by 6 percent from 2.75 in 2016 to 2.99 in 2018, with the average report quality reaching the verge of entering the upper satisfactory category in 2018. If these consistent trends in improvement continue, UNICEF should expect to cross the upper satisfactory threshold of 3.0 in 2019.

Exhibit A: Overall Evaluation Score from 2016 - 2018
As shown in Exhibit B, the percentage of evaluation reports per year rated as satisfactory or higher increased by nearly 14 percent from 73.5 percent in 2016 to 87.4 percent in 2018. At the same time, the percentage of reports rated as fair decreased by approximately half from 25.5 percent in 2016 to 12.6 percent in 2018. While 2016 produced one unsatisfactory report, no evaluations were rated as unsatisfactory in 2017 or 2018. The percentage of evaluations in the upper satisfactory and upper fair categories has also increased across the years. These trends represent overall continuous improvement in UNICEF evaluation report quality, therefore increasing the reliability of evaluative evidence for UNICEF decision-making.

**Exhibit B: Percentage of Reports per Assessment Rating from 2016 – 2018**

<table>
<thead>
<tr>
<th></th>
<th>2016</th>
<th>2017</th>
<th>2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsatisfactory</td>
<td>7.8%</td>
<td>6.7%</td>
<td>9.7%</td>
</tr>
<tr>
<td>Lower Fair</td>
<td>17.6%</td>
<td>21.3%</td>
<td>33.0%</td>
</tr>
<tr>
<td>Upper Fair</td>
<td>39.2%</td>
<td>38.2%</td>
<td>41.7%</td>
</tr>
<tr>
<td>Lower Satisfactory</td>
<td>28.4%</td>
<td>19.1%</td>
<td>12.6%</td>
</tr>
<tr>
<td>Upper Satisfactory</td>
<td>5.9%</td>
<td>14.6%</td>
<td></td>
</tr>
<tr>
<td>Highly Satisfactory</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In 2018, the average evaluation report score in each region was above the satisfactory threshold of 2.5. As seen in Exhibit C, the East Asia and Pacific Region (EAPR) produced the highest quality evaluation reports in 2018 with an average score of 3.36 followed by the Europe and Central Asia Region (ECAR) with an average score of 3.13. The Eastern and Southern Africa Region (ESAR) and the Regional Office of South Asia (ROSA) were the regions with the lowest 2018 scores with a score of 2.76 and 2.73 respectively.

A three-year cross annual trend analysis confirms that the quality of evaluation reports has been increasing overall between 2016 and 2018 for all regions with the exception of ROSA. ESAR experienced the most significant overall improvement with an increase in score of 9.3 percent followed by EAPR with an increase in score of 8.3 percent.
When separating reports by their SPOA Correspondence, evaluation reports across thematic areas are largely the same quality with only a 4.8 percent difference between the area that scored the highest (social inclusion with a score of 3.08) and lowest (WASH with a score of 2.89) in 2018. Evaluations focusing on cross-cutting priorities (including gender equality and humanitarian action) scored quite high at 3.02 and 3.06 respectively. This suggests that the GEROS rating system does not disadvantage such evaluations. In fact, humanitarian action evaluations scored second highest in 2018.

Exhibit D: Overall Evaluation Score from 2016 – 2018 by Region
A trend analysis of assessment criteria indicates that all criteria have experienced an overall improvement in quality between 2016 and 2018, as seen in Exhibit D. As of 2018, each assessment criteria had an average score of satisfactory or above. Criterion relating to Evaluation Purpose, Objectives and Scope scored highest in 2018 at 3.23 points followed by the Executive Summary (3.18 points) and the Background (3.17 points).

The use of robust evaluation methods and the integration of cross-cutting evaluation principles (human rights, equity, gender equality, and stakeholder participation) remain the weakest areas and suggest the need for continued improvement. Nevertheless, between 2016 and 2018, the criterion that experienced the most significant improvement was Evaluation Principles that improved by 12.5 percent (with a score of 2.23 in 2016 and 2.73 in 2018), bringing this criterion into the satisfactory category in 2018.
Exhibit D: Overall Evaluation Report Quality per Assessment Criterion from 2016 – 2018

**SWAP Assessment**

The overall quality of gender equality integration into evaluation reports (as per SWAP standards) has increased overall by 15 percent from 50 percent in 2016 to 65 percent in 2018. All three micro-criteria also experienced an overall increase in quality between 2016 and 2018. This being said, reports remain generally weak at using a gender responsive methodology.

All regions experienced an overall improvement in their overall SWAP score between 2016 and 2018 with the exception of ROSA that experienced a decrease in quality by 12

---

3 Micro-criteria includes three assessment questions: 1) GEEW is integrated in the evaluation scope of analysis and evaluation criteria and questions are designed in a way that ensures GEEW related data will be collected; 2) A gender responsive methodology, methods and tools, and data analysis techniques are selected; and 3) The evaluation findings, conclusions and recommendations reflect a gender analysis.
percent. LACR experienced the most significant improvement of 37 percent difference between 2016 and 2018.

**Reflections from the 2016 – 2018 GEROS Review Team**

**Strengths of the GEROS Review Tool:**

- It facilitates the standardization of the quality of evaluation reports across a decentralized structure within UNICEF;
- It provides a platform to monitor trends in evaluation quality across the organisation;
- Individual GEROS reviews per evaluation report provide real-time feedback with qualitative comments and suggestions for improvement to help UNICEF evaluation managers and evaluators improve the quality of evaluations; and
- The annual Meta-Analysis informs UNICEF decision-makers in senior management as to where to make investments to improve the quality of evaluations.

**Limitations of the GEROS Review Tool:**

- GEROS is limited to an assessment of the evaluation report and is unable to capture any events or processes that are not described within the report. UNICEF evaluation managers, therefore, must ensure that all processes are clearly described.
- While rating scores are useful for monitoring trends in evaluation quality and potentially serving as motivators for offices and regions to improve their scores, they can also shift focus away from qualitative feedback and distract from the learning aspect of the GEROS reviews.
- The current GEROS tool is unable to capture the usefulness of the evaluation to those commissioning it. The tool does not contain any mechanisms to measure the extent to which the evaluation served to improve programming.
- The current GEROS tool is unable to capture whether evaluations are drawing on previous lessons learned.
- The current GEROS tool does not include customized assessment questions per evaluation type.

The 2018 meta-analysis noted improvements in areas where evaluation performance was deemed weak by previous meta-analyses (areas needing improvement included evaluation methods, lessons learned, and evaluation principles). This trend in improvement suggests that UNICEF has made investments to improve its evaluation quality in these particular areas and that these investments are paying off. This also suggests that GEROS is a useful system for highlighting areas needing improvement and for providing guidance to UNICEF decision-makers around how to facilitate improvement.

**Conclusions and Recommendations**

The 2018 GEROS Meta-Analysis produced four (4) over-arching conclusions and corresponding recommendations to support UNICEF with the continued strengthening of its evaluation function and programming through the GEROS platform. Recommendations identify target groups for action; provide a suggested implementation timeline; and outline concrete steps for action to facilitate their implementation.

**Conclusion #1 - Remaining Weaknesses:** The overall quality of UNICEF evaluations has continued to increase in 2018, with a larger percentage of evaluation reports rated as satisfactory or higher. While improvements have been seen in areas that were identified
as weak in previous meta-analyses (particularly around methods, lessons learned, and evaluation principles), these areas still remain comparatively weak and continue to require further investment in order to bring their performance up to standard.

**Recommendation #1 - Improving Remaining Weaknesses:** UNICEF should continue to invest in increasing the quality of evaluation reports in the areas of methods, lessons learned, and evaluation principles (with a focus on SWAP principles) by taking the following action:

1) The EO should support regional offices to organize training and information-sharing webinars with country offices to share knowledge and good practices within and across regions on the areas of evaluation methods, lessons learned, and SWAP principles.
2) Evaluation Managers should continue to strengthen the requirements in the ToRs for these elements and select evaluators who have proven experience at integrating evaluation principles (particularly SWAP).
3) The EO should work with programming experts to ensure that improvements around presenting lessons learned in evaluation reports are done in parallel with organisation-wide improvements around how lessons learned are captured, shared, and used within and across regions.
4) The EO should work with the GEROS evaluation team to identify how the GEROS tool could better capture information as to where bottlenecks occur when integrating SWAP principles (this is in line with Recommendation #3 from 2017 Meta-Analysis).

**Conclusion #2 - Quality and Usefulness of Evaluation Recommendations:** The quality of evaluation recommendations has not kept pace with improvements in other areas in 2018, with some regions showing marked decreasing trends in quality. Additionally, the current GEROS system is unable to capture the usefulness of evaluation recommendations to evaluation users.

**Recommendation #2 – Improving the Quality and Usefulness of Evaluation Recommendations:** UNICEF should place a targeted focus on improving the quality of evaluation recommendations by identifying more specific evaluation users and uses and should integrate mechanisms into GEROS or use a complementary system to capture information on evaluation usefulness. UNICEF should take the following action:

1) Evaluation managers should work with programming experts to clearly identify the specific users and uses of the evaluation before it is commissioned so that evaluators can generate recommendations that are targeted and useful to evaluation users.
2) The EO should commission an exercise to clearly identify what elements make recommendations useful to end-users (apart from being targeted and actionable) and to explore how recommendations could be better used by programmers.
3) The EO should provide training to evaluation managers on how good recommendations should be made, based on good practices identified in the GEROS Meta-Analysis in order to better support evaluators with the development of useful recommendations that are linked with the evaluation purpose.
4) The EO should modify the GEROS tool or set up a complementary system to capture feedback from evaluation users on the extent to which recommendations were useful. This could include tracking the percentage of recommendations that were approved through the evaluation management responses.
Conclusion #3 – Focus on Learning: While GEROS offers huge potential to facilitate institutional learning, the current system is heavily focused on rating scores. While the use of a rating score is useful for measuring changes in evaluation quality and for incentivizing improvements in evaluation performance, it carries the risk of shifting focus away from learning if not paired with learning-specific activities.

Recommendation #3 – Increasing the Focus on Learning: UNICEF should place a stronger focus on using GEROS as a learning tool and should integrate more learning-specific activities into the GEROS process by taking the following action:

1) The EO should facilitate a more regular and systemic pairing of the GEROS system with evaluation learning opportunities (i.e. workshops, webinars, the development of institutional mechanisms to collect and share lessons learned, etc.).
2) The EO should consider adopting a mechanism to facilitate self-reflection among evaluation managers once GEROS feedback has been provided. For instance, the EO could consider including a feedback form annexed to each GEROS review where the manager could outline what he/she learned through the GEROS review and how he/she plans to use this learning to improve future evaluations.
3) The EO should modify the GEROS tool to capture information on the extent to which the object of evaluation was informed by previous evaluative evidence. To do this, GEROS criteria would need to be modified to request an assessment from evaluators on the extent to which the object of evaluation used evaluative evidence to inform its design and implementation.

Conclusion #4 – Joint Evaluation Management with other UN Organisations: As part of the UN Reform, UN organisations are increasingly expected to engage in more joint programming and evaluation. However, most UNICEF evaluations continue to be managed exclusively by the organisation. Moving forward, UNICEF will likely need to conduct more joint evaluations, for which incentives and preparations for this transition will be essential.

Recommendation #4 – Preparing and Incentivizing for Joint Evaluation Management with other UN Organisations: UNICEF should prepare for a transition towards more jointly managed evaluations with other UN organisations by identifying a formal framework to guide how evaluations will be jointly managed and to incentivize UNICEF evaluation offices to engage in joint evaluations by taking the following action:

1) The EO should work with senior management to identify the organisational priorities around evaluation in a context of increased joint management including potential financial structures, evaluation functions (including the degree of independence to be afforded to the evaluation process), desired working modalities, and intended roles and responsibilities.
2) Once priorities have been identified, the EO should document them and distribute them widely throughout the organisation, and then begin to encourage more joint evaluations throughout the organisation.
3) The EO should include targets on joint evaluations within evaluation plans.
1. Introduction

This report presents a meta-analysis of the quality of the evaluation reports submitted to UNICEF’s Global Evaluation Reports Oversight System (GEROS) during 2018, and provides a comparison and trend analysis over the past three (3) years from 2016 – 2018. It provides an analysis across the three-year Limited Term Agreement (LTA) with Impact Ready Ltd, as 2019 will mark the beginning of a new LTA with a new service provider. The report synthesizes results of 104 evaluation reports from 2018; 88 evaluation reports from 2017; and 101 evaluation reports from 2016. The report presents findings at the global level, as well as trends across regions, sectors and quality assessment criteria. Regional briefs are attached as appendices that provide a summary of each region’s performance in 2018 and present data at the country level. This report contributes to a wider body of knowledge of similar GEROS meta-analysis reports produced each year since GEROS began in 2010. Data from previous years is presented where relevant.

The UNICEF Evaluation Office commissioned this report to inform key audiences, which include: The Global Evaluation Committee, UNICEF senior management (at the global, regional and country levels), the UNICEF Evaluation Office, regional evaluation officers, and evaluation managers. Its purpose is to monitor progress, analyse strengths, and identify current challenges and areas for improvement in order to contribute towards the continued strengthening of the UNICEF evaluation function and UNICEF programming.

2. Background on the UNICEF GEROS System

The Global Evaluation Reports Oversight System (GEROS) was established in 2010 to assess the quality of UNICEF evaluation reports and to further inform organizational learning. As the UNICEF evaluation system is decentralized, GEROS provides a holistic system to ensure good quality and credibility of evaluation reports across the organisation, and serves as a platform for institutional learning.

The GEROS process includes a review by an external independent team of evaluation experts of the quality of UNICEF evaluation reports (at the country, regional and global levels) and a yearly meta-analysis. UNICEF-adapted UNEG standards are used as the criteria of quality assessment (Annex II) and inter-related reliability is managed through a quality control process.

---

4 Evaluation quality is defined according to the GEROS rating assigned to each evaluation report. The GEROS rating is the result of a quality assessment based on OECD/DAC evaluation standards. Reports rated as “satisfactory or highly satisfactory” can be used with confidence to inform decision-making; those rated as fair can be used with caution as substantial improvements could be made to the report; and reports rated as “unsatisfactory” do not meet OECD/DAC evaluation standards and can therefore not be used for decision-making.
3. Purpose, Scope and Objective of GEROS

The purpose of GEROS is to support the strengthening of the evaluation function to fully meet UNEG standards, ensure accountability, and promote the use of robust evaluative evidence.

The purpose of this meta-analysis is to support the three objectives of GEROS (particularly objective 1), which are:

**Objective 1:** Support an enabling environment for senior managers and the UNICEF Executive Board to make informed decisions based on a clear understanding of the quality of evaluation evidence and usefulness of evaluation reports;

**Objective 2:** Provide feedback that leads to stronger evaluation capacity of UNICEF and its partners; and

**Objective 3:** Support UNICEF and its partners in becoming more knowledgeable about what works, where and for whom.

GEROS is underpinned by United National Evaluation Group (UNEG) norms and standards (2016), the UN System Wide Action Plan on gender equality (UN-SWAP) and other UNICEF-adapted standards, including equity and human rights-based approaches⁵. All reports and the results of their quality assessment are made available in the UNICEF Global Evaluation and Research Database (ERDB), as well as made publicly available on the UNICEF external website.

4. Methodology

This meta-analysis was conducted from May-June 2019 once all of the evaluation reports for 2018 had been assessed, submitted to and accepted by UNICEF EO. Quantitative data was compiled regarding scores for different aspects of the reports from 2016 - 2018 using Excel. The analysis was carried out across multiple axes:

- Descriptive elements of the reports (number, thematic correspondence, type of evaluation object, type of evaluation, type of evaluation management arrangement, evaluation strategy, evaluation design, evaluation level, geographic level, etc.)
- Overall trends (including cross time)
- Trends by quality assessment criteria (including across time)
- Regional trends (across time and at country levels within the Regional Briefs; and
- UN-SWAP performance and trends

A qualitative assessment was also provided for 2018 trends by quality assessment criteria to identify strengths and areas for improvement. Comments made by reviewers on each evaluation quality assessment review for 2018 were filtered according to section and then

---

⁵ These are based on the UN Evaluation Group technical note for the UN-SWAP EPI: [http://www.unevaluation.org/document/detail/1452](http://www.unevaluation.org/document/detail/1452)
synthesized to identify trends around strengths and weaknesses. In addition, the reviews were trawled to identify good evaluation practices from the reports.

Quantitative and qualitative data were then triangulated, and compared with longitudinal data on findings from previous years to map key trends and patterns.

The report also includes a section on GEROS as a Learning Tool that presents insights from the external independent evaluation team around the GEROS tool and how it can be further strengthened to better support the evaluation function and organizational learning. This analysis should be read along with the findings and conclusions of the 2019 GEROS Review that provides a more in-depth assessment of the GEROS system, including how evaluation report quality is defined and assessed.

The GEROS review process assigns an overall rating to each evaluation report on a scale from 0 – 4 and uses a ranking system to classify the quality of reports as follows: unsatisfactory (0.0 – 1.49); fair (1.5 – 2.49); satisfactory (2.5 – 3.49); and highly satisfactory (3.5 – 4.0). The fair and satisfactory categories are further broken down into lower fair (1.5 – 1.99); upper fair (2.0 – 2.49); lower satisfactory (2.5 – 2.99); and upper satisfactory (3.0 – 3.49). For this GEROS Meta-Analysis, changes in percentage score are in relation to the maximum potential score of 4/4 (100 percent)6.

Ratings for each of the nine assessment criteria (including SWAP) are also provided per report. Each criterion is assessed using specific assessment questions (see Annex II for a description of the assessment criteria and questions). Scores per assessment criterion use a 4-point scale and assessment questions within each criterion (including SWAP) use a percentage scale (100 percent being a perfect score).

**Number of Evaluation Reports Included in this Meta-Analysis**

In 2018, a total of 103 evaluation reports were classified as evaluations and reviewed as part of the GEROS process. This is the largest number of evaluations to be annually reviewed since the start of GEROS and represents an increase in the number of reports assessed in 2017 by 15 reports. (see Exhibit 1).

**Exhibit 1. Reports Included in GEROS from 2011 - 2018**

<table>
<thead>
<tr>
<th>Year</th>
<th>Reports Reviewed</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>88</td>
<td>88</td>
</tr>
<tr>
<td>2012</td>
<td>85</td>
<td>173</td>
</tr>
<tr>
<td>2013</td>
<td>96</td>
<td>269</td>
</tr>
<tr>
<td>2014</td>
<td>69</td>
<td>338</td>
</tr>
<tr>
<td>2015</td>
<td>90</td>
<td>428</td>
</tr>
<tr>
<td>2016</td>
<td>101</td>
<td>529</td>
</tr>
<tr>
<td>2017</td>
<td>88</td>
<td>617</td>
</tr>
<tr>
<td>2018</td>
<td>103</td>
<td>720</td>
</tr>
</tbody>
</table>

---

6 For example, page 9 explains that the overall report quality from 2016 – 2018 improved by 6 percent (from 2.75 to 2.99 out of a possible score of 4.00).
The number of reports submitted by each region varies, as shown in the graph below (Exhibit 2). In 2018, the greatest number of reports was submitted from the Eastern and Southern Africa Region and the West and Central Africa Region (20 reports respectively); the least number of reports was submitted by MENA Region (8 reports).

**Exhibit 2: Number of GEROS Reports per Region and HQ from 2016 - 2018**

5. Findings

2016 - 2018 Comparative Analysis

This section provides findings that have emerged from the quantitative and qualitative analysis (including triangulation) of data from the GEROS Evaluation Quality Assurance reviews from 2018 as well as a cross-annual comparison analysis from 2016 – 2018. It includes a presentation of the trends within the UNICEF Evaluation Portfolio across years and regions; a quantitative analysis of the trends in overall evaluation quality and per assessment criterion from 2016 – 2018; and a qualitative analysis of strengths and areas for improvement of the 2018 evaluation reports per assessment criterion (with additional detail provided in Annex IV). The section ends with reflections from the external independent evaluation team around the GEROS tool and how it can be further strengthened to better support the evaluation function and organizational learning.
5.1 Trends within the UNICEF Evaluation Portfolio

2014 – 2017 Strategic Plan Objective Areas (SPOA) Correspondence

As Exhibit 3 indicates, the breakdown of thematic areas within the UNICEF Portfolio has stayed relatively stable across the three-year period. In both 2017 and 2018, the largest number of evaluations focused on health while child protection was the largest focus in 2016. The proportion of evaluations dedicated to gender equality was also substantial between 2016 and 2018, with gender equality representing 14 percent of the portfolio in 2018.

Exhibit 3: SPOA Correspondence from 2016 - 2018
2018 – 2021 Strategic Plan Goals and Cross-Cutting Themes Correspondence

As UNICEF moves forward with a new 2018-2021 Strategic Plan, Exhibit 4 provides a breakdown of how evaluations in 2018 fit thematically within the new strategic plan goal areas and cross-cutting themes. The goal areas with the largest percentage of evaluations in 2018 were the goals “every child survives and thrives” and “cross-cutting priorities” (humanitarian and gender equality evaluations) representing a little under half of the portfolio at 21 percent each.

Exhibit 4: Correspondence with UNICEF 2018-2021 Strategic Plan Goals and Cross-Cutting Themes

<table>
<thead>
<tr>
<th>Goal Area</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Every child survives and thrives</td>
<td>21%</td>
</tr>
<tr>
<td>Every child learns</td>
<td>13%</td>
</tr>
<tr>
<td>Every child is protected from violence and exploitation</td>
<td>10%</td>
</tr>
<tr>
<td>Every child lives in a safe and clean environment</td>
<td>10%</td>
</tr>
<tr>
<td>Every child has an equitable chance in life</td>
<td>17%</td>
</tr>
<tr>
<td>Cross-cutting priorities</td>
<td>18%</td>
</tr>
</tbody>
</table>

Type of Object of Evaluation

As can be seen in Exhibit 5, across 2016 – 2018, programmes were the most common objects evaluated, with the percentage of programme evaluations increasing to nearly half of the overall evaluation portfolio in 2018. Projects were the next most common objects evaluated, with a surge in project evaluations in 2017. Country programmes were a large portion of the portfolio in 2016 at 18 percent but then decreased to only 3 percent in 2018.

Irregular objects of evaluation (including organizational/business units, pilots/innovation, policy/norms/standards, strategies, systems, and thematic areas) combined comprised less than half of the portfolios per year. However, between 2016 and 2018 there was an increase in systems evaluations by 2 percent, pilot/innovation evaluations by 3 percent, and strategy evaluations by 4 percent. The percentage of thematic evaluations doubled from 3 to 6 percent between 2016 and 2018. In 2018, strategy (9 percent), pilot/innovation (9 percent), and system (7 percent) evaluations were the most common irregular objects evaluated.

Exhibit 5: Evaluation Objects from 2016 – 2018
Evaluation Type

UNICEF evaluations can be summative (primarily for accountability purposes), formative (primarily for learning purposes) or both formative and summative (for a mix of accountability and learning purposes). As seen from Exhibit 6 below, the percentage of evaluations with a focus on learning (be it exclusively formative or summative and formative) has been by and large consistent throughout the 2016-2018 period, with a subtle spike in 2017 (at 76 percent, as compared to both 2016 and 2018 when learning-focused evaluations amounted to 72 percent).

Exhibit 6: Evaluation Types from 2016 – 2018
Evaluation Approach
Throughout the three-year period, a large majority (over three fourths) of evaluations have drawn on mixed quantitative and qualitative methods. As seen from Exhibit 7 below, the percentage of reports relying exclusively on qualitative methods has decreased by 13 percent from 20 percent in 2016 to 7 percent in 2018, indicating an increasing focus on including quantitative methods in evaluation approaches.

*Exhibit 7: Trends in Evaluation Approach from 2016 – 2018*
Evaluation design

The most common evaluation designs used during the three-year period were quasi-experimental, theory-based, participatory, and case study designs. Trends indicate a steady increase in the use of quasi-experimental evaluation designs by 32 percent from 6 percent in 2016 to 38 percent in 2018. Theory-based designs also saw an overall increase from 18 percent in 2016 to 24 percent in 2018. In contrary to these increasing trends, participatory designs decreased from 30 percent in 2016 to 9 percent in 2018, along with case studies that experienced a dramatic decrease between 2017 and 2018 from 25 percent to only 5 percent.

Exhibit 8: Trends in Evaluation Design from 2016 – 2018
Evaluation Level

Exhibit 9 below shows that over the three-year period an average of at least 94 percent of evaluations were targeted at the outcome or impact levels, with nearly half of all evaluations including impact level results. In fact, the percentage of evaluations focusing exclusively at the output level decreased from 6 percent in 2016 to a mere 1 percent in 2018. This reflects ambition on the part of UNICEF to understand how its outputs have contributed towards greater results over a longer period of time and/or in collaboration with partners.

Exhibit 9: Trends in evaluation level from 2016 – 2018
Geographic Level

As Exhibit 10 indicates, most (76 percent in 2016 and 82 percent in 2018) evaluations are commissioned at the country level and are national in scope. In fact, the percentage of national / country-level evaluations slightly increased by 6 percent between 2016 and 2018. The presence of multi-region/global evaluations also increased by 5 percent from 2 percent in 2016 to 7 percent in 2018.

In contrast, the percentage of multi-country evaluations decreased by 6 percent with 7 percent in 2016 and only 1 percent in 2018. The percentage of regional evaluations also slightly decreased from 3 percent in 2016 to 1 percent in 2017 but then came back to 3 percent in 2018.

Exhibit 10: Trends in geographic level from 2016 – 2018
Evaluation Management

As Exhibit 11 indicates, over three fourths of all evaluations were managed exclusively by UNICEF from 2016 – 2018, although the percentage of UNICEF-managed evaluations decreased by 7 percent from 85 percent in 2016 to 78 percent in 2018.

The percentage of evaluations jointly managed with the national government increased slightly from 5 percent in 2016 to 8 percent in 2018, as did evaluations managed with organisations outside of the UN system (from 2 percent in 2016 to 4 percent in 2018).

Of interest is that the percentage of evaluations jointly managed with other UN agencies remained the same at a mere 2 percent across 2016 – 2018. The new UN Reform called for in 2018 requests UN organisations to increasingly conduct joint programming and evaluation.

Evaluations that were externally managed or led by the national government were nearly non-existent over the three years, with no such evaluations in 2018.

Exhibit 11: Trends in evaluation management from 2016 – 2018
5.2 Overall Evaluation Report Quality from 2016 - 2018

Quantitative Overall Score
As displayed in Exhibit 12, the average quality of UNICEF evaluation reports remained within the “lower satisfactory” rating across 2016 – 2018, with consistent improvement in overall evaluation report quality by 6 percent.

In 2018, the overall quality of evaluations was on the verge of entering the “upper satisfactory” category at 2.99. If these consistent trends in improvement continue, UNICEF should expect to enter the “upper satisfactory” category in 2019.
Exhibit 13 demonstrates that the percentage of evaluation reports per year rated as satisfactory or higher increased by 13.9 percent from 73.5 percent in 2016 to 87.4 percent in 2018. At the same time, the percentage of reports rated as fair decreased by approximately half from 25.5 percent in 2016 to 12.6 percent in 2018. While 2016 produced one unsatisfactory report, no evaluations were rated as unsatisfactory in 2017 or 2018. These trends represent overall continuous improvement in UNICEF evaluation report quality. (See Annex III for the rating of each evaluation in 2018).
By breaking down the fair and satisfactory categories into lower and upper groupings, we gain further insights around trends in evaluation quality. As Exhibit 14 outlines, the percentage of reports rated as upper satisfactory consistently increased by 13.3 percent from 2016 – 2018, suggesting that satisfactory evaluations are steadily moving towards the highly satisfactory rating. In 2018, over half of all reports rated as satisfactory were in the upper satisfactory category.

The trends for the fair ratings indicate that the majority of evaluations rated as fair are in the upper fair category. The percentage of reports rated as upper fair grew from 2016 to 2017 but then decreased in 2018 as a larger percentage of reports upgraded into the satisfactory category. In 2018, 87.3 percent of evaluations were rated as satisfactory or higher, which represents an increase of 15.4 percent from 2017.
Exhibit 14: Percentage of Reports per Sub-Grouping of Assessment Scores from 2016 – 2018

<table>
<thead>
<tr>
<th>Year</th>
<th>Unsatisfactory</th>
<th>Lower Fair</th>
<th>Upper Fair</th>
<th>Lower Satisfactory</th>
<th>Upper Satisfactory</th>
<th>Highly Satisfactory</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016</td>
<td>1.0%</td>
<td>7.8%</td>
<td>17.6%</td>
<td>39.2%</td>
<td>28.4%</td>
<td>5.9%</td>
</tr>
<tr>
<td>2017</td>
<td>2.9%</td>
<td>6.7%</td>
<td>21.3%</td>
<td>38.2%</td>
<td>19.1%</td>
<td>14.6%</td>
</tr>
<tr>
<td>2018</td>
<td>9.7%</td>
<td>33.0%</td>
<td>41.7%</td>
<td>12.6%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Quantitative Overall Score by Region

When breaking down the overall report scores by region, we can see (as presented in Exhibit 15) that in 2018, the average evaluation report score in each region was above the satisfactory threshold of 2.5. The East Asia and Pacific Region (EAPR) produced the highest quality evaluation reports in 2018 with an average score of 3.36 followed by the Europe and Central Asia Region (ECAR) with an average score of 3.13. The Eastern and Southern Africa Region (ESAR) and the Regional Office of South Asia (ROSA) were the regions with the lowest 2018 scores with a score of 2.76 and 2.73 respectively.

A three-year cross annual trend analysis confirms that the quality of evaluation reports has been increasing overall between 2016 and 2018 for all regions with the exception of ROSA. ESAR experienced the most significant overall improvement with an increase in score of 9.3 percent (from 2.39 in 2016 to 2.76 in 2018) followed by EAPR with an increase in score of 8.3 percent (from 3.03 in 2016 to 3.36 in 2018).
When examining the 2016 – 2018 portfolio by GEROS rating, it can be observed (as outlined in Exhibit 16) that all regions with the exception of ECAR and LACR increased the percentage of their evaluation reports rated as satisfactory or higher between 2016 and 2018. In 2017, several regions (ECAR, HQ, LACR and WCAR) experienced a substantial drop in satisfactory reports but were able to make important gains in 2018.

HQ experienced the most significant improvement between 2017 and 2018 with a 41 percent increase in reports rated satisfactory or higher, followed by WCAR with a 31 percent increase. In fact, WCAR had the largest percentage of reports rated satisfactory or higher in 2018 at 95 percent, followed by EAPR and ECAR (both at 92 percent) and HQ (91 percent).

In 2018, ROSA had the largest percentage of reports rated as fair (22.2 percent) followed by ESAR (20 percent). However, all of the fair evaluations from ESAR fell within the upper fair category. If ESAR continues with its steady trend in improvement, it is likely to experience a significant increase in satisfactory evaluations in the near future. EAPR, ECAR, HQ, and MENA are also well placed to increase their percentage of satisfactory evaluations as all of their evaluations rated as fair in 2018 fell in the upper fair category.
Exhibit 16: Overall Evaluation Score by Rating from 2018 by Region

<table>
<thead>
<tr>
<th>Region</th>
<th>Unsatisfactory</th>
<th>Lower Fair</th>
<th>Lower Satisfactory</th>
<th>Upper Fair</th>
<th>Upper Satisfactory</th>
<th>Highly Satisfactory</th>
</tr>
</thead>
<tbody>
<tr>
<td>EAPR</td>
<td></td>
<td>8.3%</td>
<td>41.7%</td>
<td>41.7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESAR</td>
<td>20%</td>
<td>50%</td>
<td>25%</td>
<td>5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LACR</td>
<td>9.1%</td>
<td>18.2%</td>
<td>54.5%</td>
<td>9.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROSA</td>
<td>11.1%</td>
<td>44.4%</td>
<td>33.3%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECAR</td>
<td>8.3%</td>
<td>25%</td>
<td>58.3%</td>
<td>8.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HQ</td>
<td>9.1%</td>
<td>27.3%</td>
<td>54.5%</td>
<td>9.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MENA</td>
<td>12.5%</td>
<td>37.5%</td>
<td>25%</td>
<td>25%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WCAR</td>
<td>5%</td>
<td>40%</td>
<td>45%</td>
<td>10%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Quantitative Overall Score by SPOA Correspondence
When separating reports by their SPOA Correspondence, we can see that evaluation reports across thematic areas are largely the same quality with only a 4.8 percent difference between the area that scored the highest (social inclusion with a score of 3.08) and lowest (WASH with a score of 2.89) in 2018.

Of interest is that evaluations focusing on cross-cutting priorities (including gender equality and humanitarian action) scored quite high at 3.02 and 3.06 respectively. This suggests that the GEROS rating system does not disadvantage such evaluations. In fact, humanitarian action evaluations scored second highest in 2018.

**Exhibit 17: Overall Evaluation Score from 2016 – 2018 by SPOA Correspondence**

<table>
<thead>
<tr>
<th>Criteria</th>
<th>2016</th>
<th>2017</th>
<th>2018</th>
<th>Max Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health</td>
<td>2.95</td>
<td>3.04</td>
<td>2.89</td>
<td>3.08</td>
</tr>
<tr>
<td>HIV/AIDS</td>
<td>3.00</td>
<td>3.01</td>
<td>3.01</td>
<td>3.06</td>
</tr>
<tr>
<td>WASH</td>
<td>2.97</td>
<td>3.00</td>
<td>3.01</td>
<td>3.08</td>
</tr>
<tr>
<td>Nutrition</td>
<td>3.00</td>
<td>3.01</td>
<td>3.01</td>
<td>3.06</td>
</tr>
<tr>
<td>Education</td>
<td>3.00</td>
<td>3.01</td>
<td>3.01</td>
<td>3.06</td>
</tr>
<tr>
<td>Child protection</td>
<td>3.00</td>
<td>3.01</td>
<td>3.01</td>
<td>3.06</td>
</tr>
<tr>
<td>Social inclusion</td>
<td>3.00</td>
<td>3.01</td>
<td>3.01</td>
<td>3.06</td>
</tr>
<tr>
<td>Gender equality</td>
<td>3.00</td>
<td>3.01</td>
<td>3.01</td>
<td>3.06</td>
</tr>
<tr>
<td>Humanitarian...</td>
<td>3.00</td>
<td>3.01</td>
<td>3.01</td>
<td>3.06</td>
</tr>
</tbody>
</table>

**5.3 Evaluation Report Quality per Assessment Criterion from 2016 - 2018**

The following section presents trends in evaluation report quality from 2016 – 2018 per GEROS assessment criterion. It begins with an overview of all nine (9) assessment criteria and then provides a detailed discussion of each criterion (through sub-sections A – I). A qualitative assessment of strengths and weaknesses with report examples and references to best practices per criterion is also provided. SWAP ratings are presented within the criterion on evaluation principles.

**Overview**
A trend analysis of assessment criteria indicates that all criteria have experienced an overall improvement in quality between 2016 and 2018. As of 2018, each assessment criterion had an average score of satisfactory or above.

Criterion relating to Evaluation Purpose, Objectives and Scope scored highest in 2018 at 3.23 points followed by the Executive Summary (3.18 points) and the Background (3.17 points). The use of robust evaluation methods and the integration of cross-cutting evaluation principles (human rights, equity, gender equality, and stakeholder participation) remain the weakest areas and suggest the need for continued improvement. Nevertheless, between 2016 and 2018, the criterion that experienced the most significant improvement was Evaluation Principles that improved by 12.5 percent (with a score of 2.23 in 2016 and 2.73 in 2018), bringing this criterion into the satisfactory category in 2018.

**Exhibit 18: Overall Evaluation Report Quality per Assessment Criterion from 2016 – 2018**

<table>
<thead>
<tr>
<th></th>
<th>2016</th>
<th>2017</th>
<th>2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background</td>
<td>2.93</td>
<td>3.15</td>
<td>3.17</td>
</tr>
<tr>
<td>Purpose</td>
<td>2.91</td>
<td>3.20</td>
<td>3.23</td>
</tr>
<tr>
<td>Methods</td>
<td>3.15</td>
<td>2.93</td>
<td>2.80</td>
</tr>
<tr>
<td>Findings</td>
<td>2.93</td>
<td>2.83</td>
<td>2.71</td>
</tr>
<tr>
<td>Conclusions</td>
<td>2.80</td>
<td>3.02</td>
<td>3.03</td>
</tr>
<tr>
<td>Recommendations</td>
<td>2.71</td>
<td>3.09</td>
<td>3.02</td>
</tr>
<tr>
<td>Structure</td>
<td>2.97</td>
<td>3.16</td>
<td>3.18</td>
</tr>
<tr>
<td>Principles</td>
<td>3.03</td>
<td>3.08</td>
<td>3.17</td>
</tr>
<tr>
<td>Executive Summary</td>
<td>2.34</td>
<td>2.73</td>
<td>3.08</td>
</tr>
</tbody>
</table>

**Section A: Background**

The criterion on background assesses how well the evaluation report outlines the object of evaluation and the context in which it operates, and includes four assessment questions (object, context, ToC, and stakeholders).

As Exhibit 19 indicates, the overall score from 2016 - 2018 among evaluation reports for the background criterion increased by 6 percent from a score of 2.93 in 2016 to 3.17 in 2018. The year 2017 experienced a slight decrease in quality by 0.5 percent but then
increased by 6.5 percent in 2018. In 2017, reports generally struggled to clearly identify stakeholders and their contributions to the object of evaluation. In 2018, reports were stronger at identifying stakeholders but were still inconsistent at clearly identifying the in-kind and/or financial contributions of stakeholders.

**Exhibit 19: Average Evaluation Report Quality for the “Background” Assessment Criterion from 2016 – 2018**

Looking at the Background score per region, we can see that all regions have improved their average score for this criterion with the exception of ROSA, which experienced a 12.8 percent decrease in score between 2017 and 2018. In 2018, EAPR had the highest overall average for this criterion at a score of 3.58 followed by LACR at 3.36 and ECAR at 3.33. The regions that require the most improvement for this criterion are ROSA at 2.78, WCAR at 2.95 and ESAR at 3.0.

**Exhibit 20: Average Evaluation Report Quality per Region for the “Background” Assessment Criterion from 2016 – 2018**
Overall Strengths and Weaknesses of the Background Section

As part of the background criterion, Exhibit 21 indicates that evaluations overall in 2018 were strongest at describing the context in which the object of evaluation operates, followed by presenting a strong theory of change. The area that remains weakest within this criterion is the discussion around stakeholder contributions to the design and implementation of the object of evaluation.

*Exhibit 21: Average Evaluation Report Quality per Question for the “Background” Assessment Criterion from 2016 – 2018*

Strengths:
- **Descriptive Context:** Between 2016 and 2018, reports have increasingly provided more descriptive detail around the object of evaluation and its context
(including information on beneficiaries and the financial value of the object of evaluation). (See good practice examples 1.D from Cambodia and 1.F from Mozambique in annex V).

- **Strong Theory of Change:** More reports are including theories of change (ToC) (whether they are original ToCs or those re-constructed by the evaluation team) as well as improved overall quality of ToCs. Although humanitarian evaluations are not strictly required to present a logic model or ToC, those reports that describe the logic of the intervention are typically stronger than those that do not. (See good practice example 1.G from ESARO in Annex V).

**Weaknesses:**
- **Unclear Stakeholder Contributions:** While stakeholders are often identified, many evaluation reports fail to clearly outline their contributions (financial or otherwise) to the object of evaluation. (See good practice examples in Annex V 1.A from Zimbabwe, 1.B from Cambodia, and 1.C from Palestine to learn more about how to improve in this area).

See Annex IV for a more detailed list of common strengths and weaknesses of 2018 evaluation reports related to this criterion, with examples.

**Section B: Evaluation Purpose, Objectives and Scope**

The criterion on evaluation purpose, objectives and scope assesses how well the evaluation report outlines the evaluation purpose (including intended evaluation users and uses), evaluation objectives, and evaluation scope (including what elements of the object of evaluation are and are not covered by the evaluation). It includes two assessment questions (evaluation purpose and evaluation objectives & scope).

Exhibit 22 showcases a slight but steady increase of 2 percent in the overall quality of evaluation reports relating to this criterion. This increase is due to the fact that evaluations are becoming stronger at clearly identifying evaluation objectives and at outlining what is included in the evaluation scope. It is still common, however, for evaluations to not clearly outline those elements that are not covered by the evaluation scope.

*Exhibit 22: Average Evaluation Report Quality for the “Purpose, Objectives and Scope” Assessment Criterion from 2016 – 2018*
Even if there was an overall increase in score for this criterion from 2016 – 2018, the regional breakdown shows a decrease in the score for this assessment criterion in HQ, ROSA, MENA, and WCAR. Yet, in 2018, EAPR and MENA scored the highest for this criterion at 3.67 and 3.63 respectively while ESAR and ROSA scored the lowest at 2.75 and 3.0 respectively, as can be seen from Exhibit 23 below.

**Exhibit 23: Average Evaluation Report Quality per Region for the “Purpose, Objectives and Scope” Assessment Criterion from 2016 – 2018**

Overall Strengths and Weaknesses of the Purpose, Objective and Scope Section

When looking at the assessment questions within this criterion (see Exhibit 24), we see a steady decrease in the quality of the description of the evaluation purpose between 2016 and 2018. In contrast, evaluations are increasingly providing stronger descriptions of evaluation objectives and the evaluation scope.
Exhibit 24: Average Evaluation Report Quality per Question for the “Purpose, Objectives and Scope” Assessment Criterion from 2016 – 2018

**Strengths**

- **Clear Objectives**: Evaluation objectives are typically clearly stated and are aligned with those outlined in the terms of reference (ToRs).

- **Completeness of Scope**: While more and more evaluations are providing a good description of what is included in the evaluation scope in terms of thematic, geographic, and time-bound characteristics, it is still common for evaluations to overlook mentioning what elements are not included in the evaluation scope. (See good practice example 2.B from the Evaluation Office in Annex V).

**Weaknesses**

- **Specificity around Evaluation Users and Uses**: Evaluation reports tend to provide overly vague descriptions of the evaluation users and uses. Those evaluations that have scored high for this question tend to break down evaluation users into sub-groups (such as UNICEF units, government ministries, etc.) as opposed to general users (i.e. “UNICEF” or “national government”) and they tend to identify a specific evaluation use per user. This level of specificity around how the evaluation will be used is key for ensuring that recommendations are targeted and useful to the specific evaluation users. (See good practice example 2.A from Togo in Annex V to learn more about how to improve in this area).

See Annex V for a more detailed list of common strengths and weaknesses of 2018 evaluation reports related to this criterion, with examples.

**Section C: Evaluation Methods**
The criterion on evaluation methods assesses the robustness of the evaluation design and its methods as well as ethics. It includes three assessment questions (criteria, methods, and ethics).

As previously noted in this section, “evaluation methods” was the second-lowest scoring GEROS criterion (after evaluation principles) between 2016 and 2018. Since robust evaluation methods are crucial for a good quality evaluation report, this overall weakness was highlighted to UNICEF management during both the 2016 and 2017 meta-analysis reports. Fortunately, the three-year trends analysis in Exhibit 25 shows an increase of 7 percent in the overall quality of evaluation methods between 2016 and 2018.


Between 2016 and 2018, all regions experienced an overall increase in the quality of evaluation methods. In 2018, the region with the strongest overall average score was EAPR (3.33). The regions with the lowest average score were ESAR (2.70) and HQ (2.73).

Exhibit 26: Average Evaluation Report Quality per Region for the “Methods” Assessment Criterion from 2016 – 2018
Overall Strengths and Weaknesses of Evaluation Methods

The overall score of each GEROS assessment question under the methods criterion has also improved from 2016 – 2018 with the most significant improvement seen within the ethics question (an improvement of 13 percent). This improvement in describing how evaluations followed ethical principles is a main contributor to the overall improvement in score seen within the methods criterion.

**Exhibit 27: Average Evaluation Report Quality per Question for the “Methods” Assessment Criterion from 2016 – 2018**

**Strengths:**

- **Use of Evaluation Criteria:** Overall, most evaluations are grounded in an assessment criteria model presented in an Evaluation Matrix, although not all matrices include detailed assessment indicators. (See good practice example 3.E from Benin in Annex V).
Weaknesses:

- **Justification of Evaluation Design and Methods:** While most evaluations appear to use appropriate evaluation designs and methods, reports often do not adequately explain why these designs and methods are well suited for meeting the evaluation purpose and objectives. (See good practice examples 3.B from the Evaluation Office and 3.C from Cambodia in Annex V to learn more about how to improve in this area).

- **Mitigation Strategies for Limitations:** Evaluation limitations are usually identified but not all reports provide mitigation strategies to address the limitations. (See good practice example 3.F from Palestine in Annex V to learn more about how to improve in this area).

- **Ethics:** Even though reports are becoming stronger at describing how the evaluation followed ethical principles, it is common for evaluation reports to inadequately describe the ethical obligations of the evaluators. (See good practice example 3.H from Malawi in Annex V to learn more about how to improve in this area).

See Annex IV for a more detailed list of common strengths and weaknesses of 2018 evaluation reports related to this criterion, with examples.

**Section D: Evaluation Findings**

The criterion on evaluation findings assesses the extent to which findings are based on robust evidence and sound analysis, and address the evaluation objectives and scope. It includes three assessment questions (findings, analysis, and results-based management).

As Exhibit 28 states, the overall quality of evaluation findings increased steadily by 7 percent from a score of 2.65 in 2016 to 2.93 in 2018. This trend is due to an overall increase in quality within each assessment question, with the most significant improvements occurring between 2017 and 2018.

Between 2016 and 2018, most regions also experienced an increase in the overall quality of evaluation findings, with the exception of ECAR, HQ and WCAR. These three regions experienced a decrease in quality in 2017. Although they were able to improve their scores in 2018, they were unable to surpassing their 2016 rating.

In 2018, EAPR had the strongest overall rating for evaluation findings at 3.25, followed by MENA at 3.13. Regions with the lowest scores in 2018 were ROSA (2.89), WCAR (2.90) and LACR (2.91).

**Exhibit 29: Average Evaluation Report Quality per Region for the “Findings” Assessment Criterion from 2016 – 2018**

Overall Strengths and Weaknesses of Evaluation Findings
The average scores for each assessment question also steadily increased between 2016 and 2018, as seen in Exhibit 30. In 2018, the findings question (that looked at the sufficiency of evidence and reference to the object’s results framework) had the strongest overall score (85 percent) followed by the quality of analysis (83 percent) and the assessment of the object’s results-based management system (80 percent).

Exhibit 30: Average Evaluation Report Quality per Question for the “Findings” Assessment Criterion from 2016 – 2018

Strengths:
- **Overall Reliability of Evidence:** Findings were generally based on robust evidence and presented according to the object’s results framework. (See good practice examples 4.A from Malawi and 4.B from Palestine in Annex V).
- **Use of Benchmarks:** Evaluations generally drew on adequate assessment frameworks that included benchmarks to compare progress over time. (See good practice examples 4.D from the Evaluation Office and 4.E from Lesotho in Annex V).
- **Analysis of Causal Reasons for Results:** Findings generally identified the causal reasons for results by exploring factors that supported or hindered the achievement of results. (See good practice example 4.F from Benin in Annex V).

Weaknesses:
- **Evidence at the outcome level:** While ample evidence was typically provided around results outputs, most evaluations struggled to present reliable and robust evidence at the outcome level.
- **Unexpected Findings:** While evaluations generally discussed both the strengths and weaknesses of the object of evaluation, it was common for evaluations not to adequately discuss positive and negative unintended effects of the evaluated object.
• **Assessment of the Object's RBM System:** While an increasing number of evaluations included an analysis around the object’s results-based management (RBM) system, the quality of the analysis fluctuated greatly and the analysis often failed to assess how the system was used to inform decision-making. (See good practice example 4.G from Zimbabwe in Annex V to learn more about how to improve in this area).

See Annex IV for a more detailed list of common strengths and weaknesses of 2018 evaluation reports related to this criterion, with examples.

**Section E: Evaluation Conclusions and Lessons Learned**

The criterion on evaluation conclusions and lessons learned assesses the extent to which conclusions are balanced and provide a deeper analysis around the future implications of the findings; and the extent to which lessons learned are properly identified and can be applied to other contexts. It includes two assessment questions (conclusions and lessons learned).

As Exhibit 31 outlines, the overall score for the conclusions and lessons learned criterion increased by 7.5 percent between 2016 and 2018. The overall quality slightly decreased by 0.5 percent in 2017 (due to a slight decrease in the quality of conclusions) but then increased by 8 percent in 2018. This increase in overall score is due to gradually improving lessons learned, particularly between 2017 and 2018. While lessons learned are increasingly included in the evaluations, there remains room for improvement in terms of their usability.


<table>
<thead>
<tr>
<th>Year</th>
<th>Average Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016</td>
<td>2.73</td>
</tr>
<tr>
<td>2017</td>
<td>2.71</td>
</tr>
<tr>
<td>2018</td>
<td>3.03</td>
</tr>
</tbody>
</table>

A regional analysis indicates (in Exhibit 32) that the average score for this criterion increased in all regions from 2016 – 2018 with the exception of LACR where the average
score remained the same. The most significant improvements were seen in ESAR (with a 12.3 percent increase) and WCAR (with a 10 percent increase) between 2016 and 2018.

In 2018, EAPR had the highest average score at 3.42 followed by ECAR and WCAR both at 3.25. Despite significant improvement, ESAR was the region with the lowest score (2.65) in 2018.

**Exhibit 32: Average Evaluation Report Quality per Region for the “Conclusions and Lessons Learned” Assessment Criterion from 2016 – 2018**

<table>
<thead>
<tr>
<th>Region</th>
<th>2016</th>
<th>2017</th>
<th>2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>EAPR</td>
<td>3.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECAR</td>
<td>3.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESAR</td>
<td>2.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HQ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LACR</td>
<td>2.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MENA</td>
<td>3.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROSA</td>
<td>3.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WCAR</td>
<td>2.89</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Overall Strengths and Weaknesses of Evaluation Conclusions and Lessons Learned**

Exhibit 33 outlines how the quality of evaluation conclusions has been largely stagnant since 2016 with a small decrease in quality (by 2 percent) in 2017. In contrast, the quality of lessons learned has improved continually by 15 percent (from 57 percent in 2016 to 72 percent in 2018). This improvement in lessons learned is a positive trend, especially since lessons learned were identified in the 2016 and 2017 meta-analyses as areas requiring improvement. While more reports are including lessons learned and are presenting them in a more generalizable fashion, improvements still remain to be made around presenting lessons learned that can be easily applied to other contexts and are highly usable to inform future programming.

**Exhibit 33: Average Evaluation Report Quality per Question for the “Conclusions and Lessons Learned” Assessment Criterion from 2016 – 2018**
Strengths:

- **Balanced Conclusions**: Evaluations tend to present conclusions that provide a balanced summary of the strengths and weaknesses.

Weaknesses:

- **Analytical Depth of Conclusions**: Conclusions often do not provide a sufficiently in-depth analytical discussion around the foreseeable future implications of the findings. (See good practice example 5.A from Uzbekistan in Annex V to learn more about how to improve in this area).

- **Clarity and Usability of Lessons Learned**: In 2018, most reports included lessons learned; however, they are often mixed into the findings or conclusions and cannot be easily identified or extracted to inform future programming. Additionally, lessons learned often draw primarily from what has worked well without capturing important lessons generated by elements that did not go as planned. Further improvement is still required to correctly identify lessons learned and present them in a way that makes them easy to use. (See good practice examples 5.B from Cameroon and 5.C from Barbados in Annex V to learn more about how to improve in this area).

See Annex IV for a more detailed list of common strengths and weaknesses of 2018 evaluation reports related to this criterion, with examples.

**Section F: Recommendations**

The criterion on evaluation recommendations assesses the extent to which the recommendations are grounded in the evaluative evidence and are actionable and useful to evaluation users. It includes two assessment questions (analysis and presentation).

As outlined in Exhibit 34, the overall quality of evaluation recommendations has steadily, although only slightly, improved by 1.3 percent from 2016 – 2018. This minimal overall
increase is due to the fact that half of the regions experienced a decrease in recommendations quality between 2016 and 2018.

**Exhibit 34: Average Evaluation Report Quality for the “Recommendations” Assessment Criterion from 2016 – 2018**

![Bar chart showing average evaluation report quality from 2016 to 2018]

The regional analysis presents a less consistent trend around improvement in quality. While some regions have experienced important improvement (such as EAPR at 12.5 percent and WCAR at 7 percent between 2016 and 2018), half of the regions experienced a decrease in overall quality. In fact, the overall quality of recommendations decreased by 12.3 percent in ROSA, 10.3 percent in LACR, 3.5 percent in MENA and 1.3 percent in ECAR.

In 2018, EAPR had the highest overall score at 3.5 followed by WCAR at 3.2, while ESAR and ROSA scored the lowest at 2.6 and 2.78 respectively.
Overall Strengths and Weaknesses of Evaluation Recommendations

Exhibit 36 indicates that there has been a slight overall improvement per assessment question between 2016 and 2018. The overall quality of the analysis supporting the recommendations (including the extent to which they are based on evaluative evidence, their usefulness to evaluation users in terms of implementation practicalities, and the degree of stakeholder participation in the development and/or validation of recommendations) increased slightly by 3 percent between 2016 and 2018. The quality of the presentation of recommendations (including clear identification of the target group for action and prioritization) also improved by 3 percent.

Exhibit 35: Average Evaluation Report Quality per Region for the “Recommendations” Assessment Criterion from 2016 – 2018

Exhibit 36: Average Evaluation Report Quality per Question for the “Recommendations” Assessment Criterion from 2016 – 2018
Strengths:

- **Grounded in the Evaluation Findings and Conclusions:** Most recommendations are properly grounded in the evaluative evidence and assessment presented in the findings and conclusions. (See good practice example 6.A from Mauritania in Annex V).

- **Prioritization:** Recommendations have become increasingly prioritized and are often presented in numerical order to indicate which ones should be implemented first.

Weaknesses:

- **Lack of Detail:** Many recommendations are too general to be highly useful to evaluation users. They often lack detail around how they can be implemented and which specific stakeholder(s) should be responsible for their implementation. (See good practice example 6.A from Mauritania in Annex V to learn more about how to improve in this area).

- **Extent of Stakeholder Participation:** Many recommendations continue to lack information around the extent to which stakeholders were involved in their development and/or validation. Stakeholder engagement in developing recommendations is key to ensuring that recommendations are useful to and actionable for evaluation users.

See Annex IV for a more detailed list of common strengths and weaknesses of 2018 evaluation reports related to this criterion, with examples.

**Section G: Evaluation Structure and Presentation**

The criterion on evaluation structure and presentation assesses the extent to which the report contains all necessary information within the opening pages and annexes and the extent to which it is logically structured and easy to navigate. It includes two assessment questions (completeness and structure).

As Exhibit 37 demonstrates, the overall quality of evaluation report structure and presentation improved marginally by 1.3 percent from 2016 to 2018. The year 2017 saw a larger improvement of 3.3 percent. However, the offices struggled to maintain this increase in quality in 2018 when the score decreased by 2 percent to 3.08. The decrease in score is due largely to an excessive length of several evaluations in 2018 that reduced their overall accessibility and usability.
The quality of the structure and presentation of evaluation reports does not provide clear trends across regions from 2016 – 2018. While some regions experienced an overall increase in quality (WCAR at 1.3 percent, EAPR by 2 percent, and ESAR by an impressive 10.3 percent), the average score decreased in over half of the regions. The most significant decrease in score was seen at HQ by 12 percent from a score of 3.38 in 2016 to 2.91 in 2018.

The region with the highest score in 2018 was EAPR with an average score of 3.75 and the region with the lowest score was LACR with an average score of 2.82. These variances in scores suggest that learning across regions through the sharing of best practices may be useful.

Exhibit 38: Average Evaluation Report Quality per Region for the “Structure and Presentation” Assessment Criterion from 2016 – 2018
Overall Strengths and Weaknesses of Evaluation Report Structure and Presentation

The extent to which evaluation reports were complete and included all relevant information within its opening pages and annexes improved slightly between 2016 and 2017 (by 2 percent). However, the score in 2018 reverted back to same score seen in 2016 (81 percent).

While the quality of the structure of the evaluation report (including its logical layout and accessibility to stakeholders) increased by 5 percent from 87 percent in 2016 to 92 percent in 2018, the score decreased substantially by 8 percent in 2018.

Exhibit 39: Average Evaluation Report Quality per Question for the “Structure and Presentation” Assessment Criterion from 2016 – 2018

<table>
<thead>
<tr>
<th>Score</th>
<th>Completeness</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016</td>
<td>81%</td>
<td>87%</td>
</tr>
<tr>
<td>2017</td>
<td>83%</td>
<td>92%</td>
</tr>
<tr>
<td>2018</td>
<td>81%</td>
<td>84%</td>
</tr>
</tbody>
</table>

Strengths:
- **Complete Information within the Opening Pages and Annexes**: The vast majority of evaluations presented all of the relevant information in the opening pages of the report. Annexes were also generally complete and included important information such as the Evaluation Matrix, copies of the assessment tools, the ToRs, etc.

Weaknesses:
- **Report Length**: Several evaluations in 2018 had an excessive report length (at 100 pages or over, excluding annexes), which jeopardizes their accessibility to evaluation users. Reports should be concise at no more than 60 pages to be accessible to evaluation users. Information beyond this length can be included in the annexes.

- **Structure around Evaluation Criteria**: Some reports were not structured around the evaluation criteria, making it difficult to understand how UNICEF has performed in relation to each criterion.
(See Annex IV for a more detailed list of common strengths and weaknesses of 2018 evaluation reports related to this criterion, with examples).

**Section H: Evaluation Principles**

The criterion on evaluation principles assesses the extent to which evaluation reports effectively integrate a human rights-based approach to programming (HRBAP) including equity, gender equality (including SWAP principles) and stakeholder participation. It includes three assessment questions (HRBAP in evaluation design and style; HRBAP in evaluation assessment; and SWAP principles). This section includes a separate subsection on the three SWAP criteria.

As mentioned previously, the evaluation principles assessment criterion has the lowest overall score of any of the GEROS criteria across 2016 to 2018. The need to focus on improving the integration of HRBAP into evaluation reports was highlighted in the 2016 and 2017 meta-analyses. Fortunately, there has been a continuous increase in overall score for this criterion across 2016 – 2018 by 12.5 percent going from a score of 2.23 in 2016 to 2.73 in 2018. Increases were observed in all three assessment criteria.


At the regional level, all regions have experienced an overall increase in quality regarding their integration of HRBAP with the exception of ROSA, where the quality decreased by 4.5 percent from 2.40 in 2016 to 2.22 in 2018. The most significant increases in quality were seen in LACR (which improved by a remarkable 26.3 percent from 1.77 in 2016 to 2.82 in 2018) followed by ESAR (at an impressive 15.8 percent increase from 1.92 in 2016 to 2.55 in 2018) and WCAR (at an impressive 15 percent increase from 2.15 in 2016 to 2.75 in 2018).

In 2018, the highest scoring region was HQ at an average score of 3.00 (after experiencing a 25 percent increase from 2017 to 2018). The lowest scoring region was ROSA at a score of 2.22. These impressive increases in score suggest that targeted investments by
UNICEF in evaluation report quality in response to recommendations from previous GEROS meta-analyses have paid off.

*Exhibit 41: Average Evaluation Report Quality per Region for the “Evaluation Principles” Assessment Criterion from 2016 – 2018*

Overall Strengths and Weaknesses of Evaluation Principles

Overall improvements per assessment question are also visible across 2016 – 2018 with a 12 percent increase in the use of HRBAP (including human rights, equity, gender equality, and stakeholder participation) in the design and style of evaluations, an 18 percent increase in the inclusion of HRBAP principles (including human rights, equity, gender equality, and stakeholder participation) within the evaluation assessment, and a 15 percent increase in the inclusion of gender equality as a cross-cutting theme in both the evaluation design and assessment (as per SWAP principles).

![Graph showing evaluation report quality per question for the “Evaluation Principles” Assessment Criterion from 2016 – 2018.](image)

**SWAP Assessment**

The GEROS SWAP assessment question presented above consists of three micro-criteria, which are:

1) GEEW is integrated in the evaluation scope of analysis and evaluation criteria and questions are designed in a way that ensures GEEW related data will be collected;

2) A gender responsive methodology, methods and tools, and data analysis techniques are selected; and

3) The evaluation findings, conclusions and recommendations reflect a gender analysis.

In 2016 and 2017, the SWAP assessment criteria consisted of 4 micro-criteria. In 2018, micro-criteria #1 and 2 were merged to produce the three criteria listed above. The three-year comparison from 2016 – 2018 therefore also merged the SWAP scores for micro-criteria #1 and #2 in 2016 and 2017 to ensure a standard comparison based on three criteria across the years.

Each SWAP assessment question is rated on a scale from 1 – 3, as follows:

1 = Partially Integrated
2 = Satisfactorily Integrated
3 = Fully Integrated

The overall SWAP score (comprising of the three assessment questions) is therefore based on a nine-point scale, with the following categories

0 – 3 points = Missing requirement (0% – 33.33%)
4 – 6 points = Approaches requirement (33.34% – 66.66%)
7 – 9 points = Meets Requirements (66.67% - 100%)

As Exhibit 43 points out, UNICEF evaluation reports overall were approaching SWAP requirements across 2016 – 2018, with an improvement of 15 percent from 50 percent in 2016 to 65 percent in 2018, where it overall nearly met SWAP requirements. If this improvement continues, UNICEF should expect to meet SWAP requirements overall in the near future. Evaluations are increasingly including gender specific questions in the evaluation matrix to ensure that GEEW data is collected and data collection techniques are generally designed to collect sex-disaggregated data. Evaluations are also making a more concerted effort to include some degree of gender analysis in the findings and conclusions.

All three micro-criteria also experienced an overall increase in quality between 2016 and 2018. In 2018, evaluation reports overall satisfactorily integrated SWAP principles into the evaluation scope of analysis to ensure that GEEW related data would be collected (criterion 1 at a score of 2.13). However, evaluations in 2018 only partially integrated SWAP principles in the evaluation methodology (criterion 2 at 1.82) and in the gender analysis within the evaluation findings, conclusions and recommendations (criterion 3 at 1.94).

Reports were weakest at using a gender responsive methodology as many evaluation reports were designed to include sex-disaggregated data and provide a degree of gender analysis but were not designed to ensure that women and men could participate equally in the evaluation process. Additionally, most evaluations were not designed to contribute towards advancing GEEW by selecting methods that contribute towards challenging traditional power structures. Gender analyses are often included in evaluations but lack the depth needed to make useful contributions towards challenging gender disparities.

Exhibit 43: Trends in Overall SWAP Scores and per Micro-Criterion from 2016 – 2018

As Exhibit 44 demonstrates, all regions experienced an improvement in their overall SWAP score between 2016 and 2018 with the exception of ROSA that experienced a
noticeable decrease in quality by 12 percent. LACR experienced the most significant improvement with a 37 percent difference between 2016 and 2018.

In 2018, all regions were, overall, approaching SWAP requirements with the exception of LACR, MENA, and WCAR who met the requirements overall.

**Exhibit 44: Trends in SWAP Scores per Region from 2016 – 2018**

![Trends in SWAP Scores per Region from 2016 – 2018](image)

Exhibit 45 provides a further regional breakdown for 2018 of SWAP scores per SWAP criterion. In 2018, MENA had the highest regional score across all three SWAP criteria, with the highest score at 2.38 for criterion 1 (integration in scope and analysis) and criterion 3 (gender analysis).

Exhibit 45 also demonstrates how scores across SWAP criteria can vary substantially within the same region. For instance, EAPR had a strong SWAP score for criterion 1 but significantly weaker scores for criteria 2 and 3. These trends suggest that more training around the specific elements that consist of the overall SWAP score is required across most regions.

**Exhibit 45: Trends in SWAP Scores per SWAP Criterion by Region in 2018**
Strengths:

- Throughout 2016 – 2018, we witnessed an increase in the number of evaluations that explicitly describe how stakeholders were involved at key moments during the evaluation process.

- Most evaluations were sensitive to equity considerations and made attempts to distinguish between different groups of vulnerable rights holders. (See good practice example 8.B from Guyana and Suriname in Annex V).

- Most evaluation matrices include a separate section with gender specific evaluation questions.

Weaknesses:

- While many evaluations adopt human rights language throughout the report, only some fully anchor the analysis within a human rights framework. (See good practice example 8.A from Uzbekistan in Annex V to learn more about how to improve in this area).

- In evaluation matrices, indicators that accompany non-gender specific evaluation questions are often not designed to collect sex-disaggregated data, especially qualitative data. (See good practice example 8.C from Cameroon in Annex V to learn more about how to improve in this area).

- While some initiatives that are gender-focused often elicit the views of direct beneficiaries (i.e., women/mothers), few evaluations are designed to gather the perspectives of men/fathers.

- Even though several reports present a gender analysis, many such analyses do not sufficiently explore key aspects of gender transformation. (See good practices example 8.D from Zimbabwe and 8.F from Madagascar in Annex V to learn more about how to improve in this area).

- The gender analysis is often reflected in conclusions, but not all evaluations provide recommendations for how to improve on gender. (See good practice example 8.G from Mauritania in Annex V to learn more about how to improve in this area).
See Annex IV for a more detailed list of common strengths and weaknesses of 2018 evaluation reports related to this criterion, with examples.

**Section I: Executive Summary**

The criterion on executive summary assesses the extent to which the executive summary can be used for decision-making, looking at whether it is the relevant conciseness and depth for primary intended users, whether it includes all necessary elements, and whether it can stand alone without adding information not already present in the rest of the report. The criterion contains only one assessment question (executive summary).

As Exhibit 46 indicates, the overall quality of the executive summary of evaluation reports slightly increased by an overall 3.8 percent from a score of 3.03 in 2016 to 3.18 in 2018, after experiencing a small dip in quality of 4 percent in 2017 due to incompleteness.

**Exhibit 46: Average Evaluation Report Quality for the “Executive Summary” Assessment Criterion from 2016 – 2018**

A regional analysis demonstrates (Exhibit 47) that half of the regions experienced an increase in overall quality of their executive summaries between 2016 and 2018 while the other half experienced a decrease. The biggest increase in quality was seen in WCAR at 10.5 percent while the biggest decrease in quality was seen in ROSA at 18.5 percent.

In 2018, EAPR had the highest overall score at 3.67 followed by WCAR at 3.50 and MENA at 3.25. In 2018, ROSA was significantly behind the rest of the regions in terms of the overall quality of its executive summaries at a score of 2.33.

**Exhibit 47: Average Evaluation Report Quality per Region for the “Executive Summary” Assessment Criterion from 2016 – 2018**
Overall Strengths and Weaknesses of the Executive Summary

Strengths:
- The overwhelming majority of evaluations present executive summaries that provide a good summary of key findings and recommendations, and do not introduce any information that has not already been introduced in the main report. (See good practice example 9.A from Cambodia in Annex V).

Weaknesses:
- While the majority of executive summaries include all the necessary elements, some omit to identify the intended audience while others do not provide a sufficient description of the object of evaluation. (See good practice example 9.A from Cambodia in Annex V to learn more about how to improve in this area).

- While the ideal length for an executive summary (i.e., 5-7 pages) is generally respected, there are cases where the executive summary is too lengthy, thereby limiting its usefulness for high-level users.

5.4 Reflections from the 2016 – 2018 GEROS Review Team

GEROS As a Learning Tool
The following section presents a series of reflections from the GEROS Review Team, which is comprised of professional independent evaluators working with the partnership Impact Ready Ltd., over the course of the current LTA period from 2016 – 2018. The section presents a discussion on: 1) the strengths and limitations of the current GEROS tool; 2) lessons emerging from the GEROS reviews and UNICEF improvements over the years; 3) how the GEROS system could be more learning-focused; and 4) how the GEROS tool could be improved.

Strengths and Limitations of the Current GEROS Tool
Strengths: GEROS is an important tool that serves four primary functions for UNICEF:

1) It facilitates the standardization of the quality of evaluation reports across a decentralized structure within UNICEF;
2) It provides a platform to monitor trends in evaluation quality across the organisation;
3) Individual GEROS reviews per evaluation report provide real-time feedback with qualitative comments and suggestions for improvement to help UNICEF evaluation managers improve the quality of evaluations; and
4) The annual Meta-Analysis informs UNICEF decision-makers in senior management as to where to make investments to improve the quality of evaluations.

Limitations: While the functions presented above are important strengths and serve to justify the presence of GEROS as an important organisational tool, GEROS inherently faces some limitations, the most notable of which are:

- GEROS is limited to an assessment of the evaluation report. The assessment system is based on a systematic review of evaluation reports in order to gain insights into the quality of the evaluation process and products. However, by basing the assessment on the quality of the evaluation report, the system is unable to capture any events or processes that were not described within the evaluation report. UNICEF evaluation managers, therefore, must ensure that all processes are clearly described.

- While rating scores are useful for monitoring trends in evaluation quality and potentially serving as motivators for offices and regions to improve their scores, they can also shift focus away from qualitative feedback and distract from the learning aspect of the GEROS reviews.

- The current GEROS tool is unable to capture the usefulness of the evaluation to those commissioning it. The tool does not contain any mechanisms to measure the extent to which the evaluation served to improve programming.

- The current GEROS tool is unable to capture whether evaluations are drawing on previous lessons learned.

- The current GEROS tool does not include customized assessment questions per evaluation type.

Lessons Emerging from the GEROS Reviews and UNICEF Improvements over the Years

This year’s 2018 meta-analysis, presenting trends from 2016 – 2018, has shown that the overall quality of UNICEF evaluations has improved over the past three years with overall improvements in nearly all regions. Having clear GEROS evaluation performance criteria to work towards has likely been a factor in facilitating improvement. Additionally, informing offices about GEROS criteria and helping them to understand the nuances of the criteria are also important facilitating factors. The production of the GEROS Handbook has likely been a useful tool in this regard.
In particular, the 2018 meta-analysis noted improvements in areas where evaluation performance was deemed weak by previous meta-analyses (areas needing improvement included evaluation methods, lessons learned, and evaluation principles). This trend in improvement suggests that UNICEF has made investments to improve its evaluation quality in these particular areas and that these investments are paying off. This also suggests that GEROS is a useful system for highlighting areas needing improvement and for providing guidance to UNICEF decision-makers around how to facilitate improvement.

How the GEROS System could be more Learning-Focused

Over the course of the three-year LTA, the GEROS Review Team has identified a number of ways in which the GEROS system could be more learning-focused. These include:

- Continuing to safeguard the independence of the GEROS review team so that reviews are independent and focused on areas needing improvement as opposed to satisfying the desire from offices to receive a favorable score;

- Integrating learning activities within the GEROS process to support offices to further understand how they can improve their evaluation performance. These could include training webinars around specific evaluation themes, performance summary briefs provided to regions during key moments of their evaluation cycle, etc.

- An informal learning-oriented response could be requested from evaluation managers by the Evaluation Office once they have received their GEROS review outlining what the evaluation managers learned through the GEROS review and what action they plan to take to improve evaluation report quality in the future. This could help place a stronger focus on the qualitative feedback and learning elements of the GEROS review as opposed to the score.

How the GEROS Tool could be Improved

The GEROS Tool has evolved along with the overall system since its establishment. While significant improvements have been made to the tool over the years, there remains room for further enhancement. UNICEF commissioned a review of the GEROS system in 2019, which will provide concrete recommendations on how to improve the tool moving forward. The following provides some contributing suggestions from the current GEROS review team around improvements that could be made to the tool:

- The GEROS template adopted a weighting system across assessment criteria in 2016, which helped place a greater emphasis on those report elements that are most crucial for ensuring a high quality evaluation report (such as the soundness of the methodology and analysis of the findings). Moving forward, the weighting system could be further fine-tuned based on the past three years of user-based experience. For instance, it would be useful to re-examine the number of assessment questions and sub-criteria per section since the total score per section is based on the average score of the questions and sub-criteria. As of now, those sections with fewer questions and sub-questions inevitably place a greater weight on each question that those sections with more questions.
• If UNICEF wishes to incentivize the commissioning of non-traditional evaluations or objects of evaluation (i.e. such as innovation or evaluations at a strategic level), the GEROS tool could potentially be adapted to award additional points to encourage such evaluations.

• An additional assessment question could be added to the tool requesting information on the extent to which the object of evaluation drew on previous evaluative evidence and lessons learned.

• An additional assessment question could be added to the tool requesting information on how the evaluation findings and recommendations are expected to contribute (in concrete terms) to improve future programming, including specific organizational learning when evaluations examine programmes that are coming to an end.

• Modifications could be made to the assessment criteria to better reflect some of the elements that are unique to irregular objects of evaluations.

• Modifications could be made to the assessment criteria to better understand where the bottlenecks are in terms of integrating evaluation principles (this was a recommendation made in the 2017 GEROS Meta-Analysis)
6. Conclusions and Recommendations

The following section presents four conclusions emerging from the findings of this year’s 2018 meta-analysis and are paired with corresponding actionable recommendations.

**Conclusion #1 - Remaining Weaknesses:** The overall quality of UNICEF evaluations has continued to increase in 2018, with a larger percentage of evaluation reports rated as satisfactory or higher. While improvements have been seen in areas that were identified as weak in previous meta-analyses (particularly around methods, lessons learned, and evaluation principles), these areas still remain comparatively weak and continue to require further investment in order to bring their performance up to standard.

In 2018, the overall quality of UNICEF evaluations continued to improve, with a growing number of reports crossing the threshold from fair to satisfactory and from satisfactory to highly satisfactory. Major areas that were identified in previous meta-analyses since 2016 as needing improvement also saw important progress. In particular, the articulation of evaluation methods, the presentation of lessons learned, and the use of evaluation principles (including the integration of human rights, equity, gender equality, and stakeholder participation) all increased in quality in 2018. This suggests that targeted investments to improve particular evaluation elements based on GEROS feedback has led to results. Even so, these areas continue to remain below the overall assessment average and would require continued further investment in order to bring them to par with the other assessment criteria.

**Recommendation #1 - Improving Remaining Weaknesses:** UNICEF should continue to invest in increasing the quality of evaluation reports in the areas of methods, lessons learned, and evaluation principles (with a focus on SWAP principles).

- **Target Group(s) for Action:** UNICEF Evaluation Office, Regional Offices, Country Offices, and Evaluation Managers
- **Priority and Timeframe:** Immediately (within the next 12 months)
- **Take action by (in order of priority):**

1) The EO should support regional offices to organize training and information-sharing webinars with country offices to share knowledge and good practices within and across regions on the areas of evaluation methods, lessons learned, and SWAP principles.

2) Evaluation Managers should continue to strengthen the requirements in the ToRs for these elements and select evaluators who have proven experience at integrating evaluation principles (particularly SWAP).

3) The EO should work with programming experts to ensure that improvements around presenting lessons learned in evaluation reports are done in parallel with organisation-wide improvements around how lessons learned are captured, shared, and used within and across regions.

4) The EO should work with the GEROS evaluation team to identify how the GEROS tool could better capture information as to where bottlenecks occur when integrating SWAP principles (this is in line with Recommendation #3 from 2017 Meta-Analysis).
Conclusion #2 - Quality and Usefulness of Evaluation Recommendations: The quality of evaluation recommendations has not kept pace with improvements in other areas in 2018, with some regions showing marked decreasing trends in quality. Additionally, the current GEROS system is unable to capture the usefulness of evaluation recommendations to evaluation users.

While the overall quality of evaluation recommendations increased slightly in 2018, there was also a decreasing trend in quality in some regions (half of the regions experienced a decrease in overall quality between 2016 and 2018). Recommendations are generally well grounded in the evaluation findings and conclusions and are typically prioritized. However, they often lack detail and are presented at a level that is too general to be highly useful to evaluation users. They are often missing information around how they can be implemented and do not clearly identify which specific stakeholder(s) should be responsible for their implementation. Evaluation reports also lack information around the extent to which stakeholders were involved in their development and/or validation.

Providing a detailed description of how recommendations can be implemented and by whom, as well as ensuring stakeholder engagement in their development are key elements to ensuring that recommendations will be useful to and actionable for evaluation users. Addressing these weaknesses will be essential for UNICEF to ensure that its evaluations are effectively used to improve programming.

Additionally, the GEROS tool currently does not include or is not accompanied by any mechanisms to gather information on the extent to which the recommendations have been useful to evaluation users, therefore limiting the tool’s ability to assess evaluation usefulness to improve UNICEF programming.

Recommendation #2 – Improving the Quality and Usefulness of Evaluation Recommendations: UNICEF should place a targeted focus on improving the quality of evaluation recommendations by identifying more specific evaluation users and uses and should integrate mechanisms into GEROS or use a complementary system to capture information on evaluation usefulness.

Target Group(s) for Action: UNICEF Evaluation Office and Evaluation Managers

Priority and Timeframe: Immediately (within the next 12 months)

Take action by (in order of priority):

1) Evaluation managers should work with programming experts to clearly identify the specific users and uses of the evaluation before it is commissioned so that evaluators can generate recommendations that are targeted and useful to evaluation users.

2) The EO should commission an exercise to clearly identify what elements make recommendations useful to end-users (apart from being targeted and actionable) and to explore how recommendations could be better used by programmers.

3) The EO should provide training to evaluation managers on how good recommendations should be made, based on good practices identified in the GEROS Meta-Analysis in order to better support evaluators with the development of useful recommendations that are linked with the evaluation purpose.

4) The EO should modify the GEROS tool or set up a complementary system to capture feedback from evaluation users on the extent to which recommendations were useful. This could include tracking the percentage of recommendations that were approved through the evaluation management responses.
Conclusion #3 – Focus on Learning: While GEROS offers huge potential to facilitate institutional learning, the current system is heavily focused on rating scores. While the use of a rating score is useful for measuring changes in evaluation quality and for incentivizing improvements in evaluation performance, it carries the risk of shifting focus away from learning if not paired with learning-specific activities.

GEROS serves two purposes: accountability and learning. While rating scores are useful for monitoring trends in evaluation quality and potentially serving as motivators for offices and regions to improve their scores, they can also shift focus away from qualitative feedback and distract from the learning aspect of the GEROS reviews. This presents a need to re-focus the system more towards its learning benefits. GEROS is a powerful tool that, when harnessed towards learning, has the potential to facilitate significant improvements in evaluation quality and the use of evaluations to improve programming.

Recommendation #3 – Increasing the Focus on Learning: UNICEF should place a stronger focus on using GEROS as a learning tool and should integrate more learning-specific activities into the GEROS process.

Target Group(s) for Action: UNICEF Evaluation Office, Regional Offices, and Country Offices
Priority and Timeframe: Medium Term (within the next 24 months)
Take action by (in order of priority):

1) The EO, ROs, and COs should facilitate more regular and systemic pairing of the GEROS system with evaluation learning opportunities (i.e. workshops, webinars, the development of institutional mechanisms to collect and share lessons learned, etc.).

2) The EO should consider adopting a mechanism to facilitate self-reflection among evaluation managers once GEROS feedback has been provided. For instance, the EO could consider including a feedback form annexed to each GEROS review where the manager could outline what he/she learned through the GEROS review and how he/she plans to use this learning to improve future evaluations.

3) The EO should modify the GEROS tool to capture information on the extent to which the object of evaluation was informed by previous evaluative evidence. To do this, GEROS criteria would need to be modified to request an assessment from evaluators on the extent to which the object of evaluation used evaluative evidence to inform its design and implementation.

Conclusion #4 – Joint Evaluation Management with other UN Organizations: As part of the UN Reform, UN organizations are increasingly expected to engage in more joint programming and evaluation. However, most UNICEF evaluations continue to be managed exclusively by the organization. Moving forward, UNICEF will likely need to conduct more joint evaluations, for which incentives and preparations for this transition will be essential.

In 2018, most evaluations were managed exclusively by UNICEF with a small percentage managed jointly with other UN organizations. The UN Reform is currently calling for UN organizations to increase their levels of collaboration and joint management modalities in all areas (including evaluation). Within this new context, UNICEF will likely be required to
shift towards commissioning more evaluations that are jointly managed with other UN organizations. The way in which evaluations are managed differs greatly among organisations. UNICEF uses a decentralized programming and evaluation system while some other organizations use a more centralized system. This will require UNICEF to have a well thought-through framework around how it wishes to approach increased joint management.

**Recommendation #4 – Preparing and Incentivizing for Joint Evaluation Management with other UN Organizations:** UNICEF should prepare for a transition towards more jointly managed evaluations with other UN organizations by identifying a formal framework to guide how evaluations will be jointly managed and to incentivize UNICEF evaluation offices to engage in joint evaluations.

- **Target Group(s) for Action:** UNICEF Evaluation Office
- **Priority and Timeframe:** Medium Term (within the next 24 months)
- **Take action by (in order of priority):**

1) The EO should work with senior management to identify the organizational priorities around evaluation in a context of increased joint management including potential financial structures, evaluation functions (including the degree of independence to be afforded to the evaluation process), desired working modalities, and intended roles and responsibilities.

2) Once priorities have been identified, the EO should document them and distribute them widely throughout the organization, and then begin to encourage more joint evaluations throughout the organization.

3) The EO should include targets on joint evaluations within evaluation plans.