UNICEF GEROS Meta-Analysis

An independent review of UNICEF evaluation report quality and trends

2016 – 2018

Principal Author: Katherine Garven
Contributing Authors: Fernando Garabito and Esther Rouleau
OVERVIEW

❖ GEROS Objectives
❖ GEROS Methodology
❖ 103 reports reviewed to UNEG standards in 2018
❖ Findings: Trends within the UNICEF Evaluation Portfolio
❖ Findings: Overall Evaluation Report Quality from 2016 – 2018
❖ Findings: Evaluation Report Quality by Region from 2016 – 2018
❖ Findings: Overall Evaluation Score per SPOA Correspondence
❖ Findings: Evaluation Report Quality per Assessment Criterion
❖ Findings: SWAP Assessment
❖ Reflections from the 2016 – 2018 GEROS Review Team
❖ Conclusions & Recommendations
❖ Thank you
Contribute to achieving the three (revised) objectives of GEROS (particularly objective 1):

Objective 1: Enabling environment for senior managers and executive board to make informed decisions based on a clear understanding of the quality of evaluation evidence and usefulness of evaluation reports;

Objective 2: Feedback leads to stronger evaluation capacity of UNICEF and partners;

Objective 3: UNICEF and partners are more knowledgeable about what works, where and for whom.
**GEROS METHODOLOGY**

**Evaluation Quality:** 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.

**Rating Scale:** 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).
103 REPORTS REVIEWED TO UNEG STANDARDS IN 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>
FINDINGS
**FINDINGS  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.
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.
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.

Trends represent overall continuous improvement in UNICEF evaluation report quality, therefore increasing the reliability of evaluative evidence for UNICEF decision-making.
In 2018, the average evaluation report score in each region was above the satisfactory threshold of 2.5.

EAPR produced the highest quality evaluation reports in 2018 while ESAR and ROSA had the lowest scores.

The quality of evaluation reports has been increasing overall between 2016 and 2018 for all regions with the exception of ROSA, with ESAR and EAPR experiencing the most significant improvements.
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 high, suggesting that the GEROS rating system does not disadvantage such evaluations.
Findings

Evaluation Report Quality per Assessment Criterion

All criteria have experienced an overall improvement in quality between 2016 and 2018, with each criterion scoring satisfactory or above in 2018.

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.
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.

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.

Criterion 1 = GEEW is integrated in the evaluation scope and analysis
Criterion 2 = Gender responsive methodology
Criterion 3 = Evaluation findings, conclusions and recommendations reflect a gender analysis.
REFLECTIONS FROM THE GEROS REVIEW TEAM
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.
CONCLUSIONS & RECOMMENDATIONS
CONCLUSIONS AND 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.
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:

- 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.

- 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).

- 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.

- 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).
CONCLUSIONS AND RECOMMENDATIONS

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:

- 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.

- 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.

- 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.

- 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:

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.).

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.

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:

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.

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.

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