UNICEF Global Evaluation Report Oversight System

Quality Review of Evaluation Reports 2010

Version 2.2

- FINAL -

Prepared for //

UNICEF Evaluation Office

Date // 2 March 2011

By// Joseph Barnes
Hatty Dinsmore
Sadie Watson

IOD PARC is the trading name of International Organisation Development Ltd//

Omega Court
362 Cemetery Road
Sheffield
S11 8FT
United Kingdom

Tel: +44 (0) 114 267 3620
www.iodparc.com

Acronyms

CCC	Core Commitments to Children
CEE/CIS	Central and Eastern Europe / Commonwealth of Independent States (Regional Office)
CFS	Child Friendly Schools
CO	Country Office
EAPRO	East Asia and Pacific Regional Office
EQOS	(UNICEF) Evaluation Quality Oversight System
ESARO	East and Southern Africa Regional Office
GEROS	(UNICEF) Global Evaluation Report Oversight System
GEWE	Gender Equality and Women's Empowerment
HRBAP	Human Rights Based Approach to Programming
IEC	Information, Education and Communication (materials)
M&E	Monitoring and Evaluation
MENA	Middle East and North Africa (Regional Office)
MfDR	Managing for Development Results
MTSP	Medium Term Strategic Plan
N/A	Not Applicable
RBM	Results Based Management
RO	Regional Office
ROSA	Regional Office of South Asia
TACRO	The Americas and Caribbean Regional Office
TOR (ToR)	Terms of Reference
UNEG	United Nations Evaluation Group
UNICEF	United Nations Children's Fund
WASH	Water, Sanitation and Hygiene
WCARO	West and Central Africa Regional Office

Executive Summary

UNICEF Evaluation Office (EO) has put in place a Global Evaluation Report Oversight System to monitor the impact of efforts to strengthen the UNICEF evaluation function globally.

The main purpose of this quality review process is to provide decision makers in UNICEF with information about evaluation reports that better supports using and improving the knowledge generated by the evaluation function. It seeks to go beyond raising awareness of quality issues, and demonstrate the implications of trends in evaluation quality on the usability of knowledge in the pursuit of delivering results for children and women.

This quality review process covered all evaluation reports submitted to the UNICEF Global Evaluation Report Oversight System for 2009. Reviews, research and other types of reports were excluded. The quality review tool assesses the evaluation report as a *standalone document*. The standards against which evaluation reports are assessed are set by the UNICEF deployment of the United Nations Evaluation Group (UNEG) global evaluation report standards.

The overall objective was to assess and rate the quality of evaluation reports commissioned by UNICEF in 2009 using the UNEG/UNICEF Evaluation Report Standards. Specific objectives included:

- Review and rate (with justifications) the quality of the main elements of evaluation reports, including structure, context, purpose, methodology, findings, conclusions, recommendation and lessons learned;
- To provide constructive feedback for evaluation commissioners to improve future evaluations;
- To provide a global analysis of key trends, strengths, weaknesses, and lessons of UNICEF evaluation reports; and
- To provide actionable conclusions and recommendations to improve the quality oversight system and the systemic quality of the evaluation function.

This meta-evaluation report draws on the complete sample frame of 96 reviews:

The Americas and Caribbean	12	Middle East and North Africa	6
Central and Eastern Europe, Commonwealth of Independent States	16	South Asia	11
East Asia and the Pacific	17	West and Central Africa	12
East and Southern Africa	19	Headquarters Divisions	3

Evaluation reports were initially classified according to the UNICEF evaluation typology. This allowed analysis according to various evaluation report characteristics. Each review was undertaken by an evaluation expert familiar with previous meta-evaluations of UNICEF evaluation report quality. All reviewers participated in a co-design workshop that enabled a common understanding of the standards to be reached. Three additional levels of quality assurance were applied.

The review tool primarily adopted a qualitative approach to rating evaluation reports against the overall standard of confidence. It pursued a systematic process of aggregating qualitative ratings of 58 guiding questions about different aspects of an evaluation report into six sections, and then into a final overall assessment. The six sections of the review tool were: 1/ Object of the evaluation; 2/ Purpose, objectives and scope; 3/ Evaluation methodology, gender, human rights and equity; 4/ Findings and conclusions; 5/ Recommendations and lessons learned; 6/ Report is well structured, logical and clear; 7/ (Plus additional information and an overall reaction).

This qualitative approach was designed to enable reviewers to provide useful analysis across the range of evaluation contexts encountered; and constructive feedback to improve future evaluation reports. Each question, each section and the overall report are given a rating of either 'very confident', 'confident', 'almost confident' and 'no confidence' (where relevant, a N/A option was also provided). Each rating was informed by three factors: a prompting question in the review tool; a 'confidence-to-act' test, and any ratings in the level of analysis below.

In addition to ratings, commentary was provided against each rating, suggestions for future improvement provided for each section, and executive feedback provided for each section and the overall report. The complete review process generates three types of data: a report typology, a series of ratings, and a structured set of discussion text.

The review process generated an extensive dataset to inform the trend analysis, consisting of 1,152 individual pieces of evaluation typology data, 6,432 individual ratings, and 6,912 sections of qualitative text (approximately 140,000 words). In order to distil the key findings from this data, a multi-stage process was adopted using qualitative analysis tools such as inductive coding.

The review process itself was subject to the limitation of only having access to the written evaluation report. As a direct consequence of this, the findings and conclusions drawn can only be applied to an evaluation report, and not to the evaluation itself. Qualitative analysis (as with all analysis) requires for judgements to be made in identifying the important indicators and trends contained within the dataset generated by the review process.

Findings

The meta-evaluation found that 36% of reviewed evaluation reports met the UNICEF standards to a degree that could be considered satisfactory. Whilst the remaining 64% of reports were rated as unsatisfactory, the vast majority of these (exactly half of all reports) could have been improved to a satisfactory level with just a little more work.

Overall, four evaluation reports were flagged as outstanding best practice, although six more achieved a very confident rating in one or more of the review sections. The outstanding evaluations were Thailand's Evaluation of Children and the 2004 Indian Ocean Tsunami, Guinea Conakry/Guinea Bissau's Evaluation of WASH Activities, Timor Leste's Evaluation of the UNICEF Education Programme, and Uzbekistan's Evaluation of the Family Education Project.

Reviewers noted in particular that unclear objectives were a major contributor to poor report quality, as was having an unclear purpose, inadequate evaluation questions, or missing evaluation criteria. In nearly all of these cases it was found that either poor quality terms of reference contributed to the weaknesses evident in evaluation reports.

Output level reports (30% of all reports) displayed concerning weakness across all review sections, such that 90% were rated as unsatisfactory. Outcome and impact reports were consistent across both result levels, with around 50% being satisfactory overall. Only impact reports included those that were considered to be outstanding best practice. Up to three times more summative reports than formative reports were classified as satisfactory in various aspects of the review.

Cross-referencing ratings with MTSP-correspondence reveals a consistent story. Multi-sector and cross-cutting evaluations register strongly in all sections, with around half of these reports being rated as satisfactory. Conversely, organisational performance evaluations were continuously ranked as unsatisfactory across all review sections (with the notable exception of the Global Evaluation of DevInfo).

From among the MTSP focus areas, *young child survival and development* suffered from particularly weak evaluation reports, less than 20% of which were considered to be satisfactory (albeit two being

rated as outstanding). *Policy and advocacy*, and *child protection* were fairly robust in terms of 'description of the object' and 'purpose' sections, but rated very poorly in all other sections. It was concerning to note that not a single policy evaluation report was rated satisfactory in relation to 'recommendations and lessons learned'.

HIV/AIDS evaluation reports were assessed to be poor in relation to all sections of the review, in particular with regard to methodological rigour. At the other end of the scale, *basic education and gender* evaluation reports were found to be consistently the strongest of all the MTSP focus areas with more than half of reports rated as satisfactory in most sections of the review.

Around two-thirds of evaluation reports fail to explicitly articulate the results chain of the evaluated object. Over half of reports were still able to present a satisfactory context through inclusion of other information, but the consequence of this issue is that a majority of reports do not appear to be guided by the logic of the programme or project being evaluated. An observation by reviewers accounting over half of the reports was that there is a tendency to provide general information about a country or the implementation context of the evaluated object, rather than analysis that can shape the evaluation purpose, objectives, and findings.

Reviews of the purpose, objectives and scope of evaluation reports revealed a number of underlying issues with the framing of evaluations. These appear to manifest themselves through weak justification of evaluation criteria, and lack of consistent use of these criteria within evaluations. Despite these challenges, the purpose sections of more than half of reports still manage to rate as satisfactory. Four *outstanding* reports were noted primarily for having very strong evaluation frameworks. These clearly referenced the OECD DAC evaluation criteria in addition to identifying and integrating relevant rights instruments, such as the Core Commitments to Children.

A third of reviewers found methodologies to be narrow and inadequately explained as a general rule. This is manifested in terms of weak control of bias: with a handful of reports doing no more than the evaluator providing personal reflections on a narrow set of interviews. 83% of reports do not include any discussion on ethics, although it is sometimes evident from the approaches adopted by evaluators that ethical considerations had been borne in mind at some point.

Around 10% of evaluation reports were praised for collecting diverse datasets and large potential bodies of evidence. The best of these were able to convert this data systematically into evidence, findings and conclusions using strong and transparent analysis. The majority of reports were unable to demonstrate this systematic use of evidence to construct robust findings and conclusions. Indeed, there appears to be a persistent problem in regard to data analysis, with reports not using data to its full potential.

Reviewers noted that it was often hard to see the link between recommendations and the preceding findings and conclusions. Lessons learned proved to be even more problematic than recommendations. When they were found in reports, lessons learned were more-often-than-not found to be project or programme-specific observations and not generally applicable to other contexts.

All of the seven policy evaluations included in the review were rated as unsatisfactory in relation to recommendations and lessons learned. This is an issue of some significance in relation to UNICEF's commitment to more upstream working.

The majority of evaluation teams do not appear to have had sight of the UNICEF/UNEG minimum standards. In the 34% of reports that did have TOR attached it was largely found that TORs did not draw attention to these standards.

Although 86% of evaluations did include an executive summary, these were largely found to be weak and not fit for purpose. Only 30% of reports included executive summaries that could confidently be used for decision making purposes.

The review tool proved to be challenging in terms of drawing out lessons about gender, equity and human rights. Just under half of evaluation reports (40) integrated *gender* considerations to some degree. Only seven reports dealt substantively with issues of *equity* and only 30% reports were found to have methodologies that were appropriate for analysing gender and *human rights* issues identified in their scope. One set of evaluations that did stand out as being both strong on rights and strong overall were the various Child Friendly School evaluations.

Whilst there were inevitable disparities across the quality of reports submitted by different regions, nearly all regions had at least one evaluation report that was rated *outstanding* in one or more of the assessed sections. The highest levels of satisfactory reports were concentrated in the Asian and European-based regional offices. It must also be recognised that the three global-level evaluations conducted by HQ Divisions were consistently rated as *confident* across all sections and overall.

During the review process, a number of notable practices were observed in each of the UNICEF regions. These were noted in the final section of the review tool and reported back to evaluation managers.

Conclusions

Conclusions were developed by analysing the findings for trends in underlying factors that contributed to the performance of evaluation reports

Evaluation reports benefit from having access to relevant and well-developed international frameworks. The *MTSP* and *purpose* analyses clearly reveal a tendency for evaluations of education and humanitarian objects to have reports of better quality than their contemporaries. Our investigation into this trend suggests that these two areas benefit from having well-known, mature and contextually-adaptable frameworks.

A disjuncture exists between successful evaluation and strong integration of rights. There would seem to be a complex and multi-faceted dynamic around an apparent disparity between reports that respond well to rights issues and reports that rate well overall. From the evidence available to this review, it would appear that fragmentation of rights skills from evaluation skills and unmet needs for strong mainstreaming frameworks are two central drivers to this whole dynamic.

The evaluation function is not delivering consistent contributions to upstream knowledge management. The *purpose* and *MTSP* analyses both found that policy evaluation reports and organisational performance evaluation reports are weak areas. From the perspective of UNICEF's upstream ambitions, the current performance of the evaluation function is likely to be of some concern.

Robust and transparent analysis of data is a problem. All the different ways of breaking down the rating data reveal one consistent trend: evaluation reports are stronger in the initial sections of the review, with performance gradually deteriorating over the span of the report. The central issue appears to be that evaluators are far clearer about the theory of evaluation (purpose, objectives, methodology, data collection) than the processing and analysis of data that is generated.

Weak terms of reference are contributing to poor report quality. Reports tended to 'build' off of the TOR as a starting point in terms of the evaluation purpose and framework, so better TORs inevitably resulted in better reports. This reemphasises the value of conducting basic checks and quality assurance on TORs, ensuring that each TOR gives evaluators sight of the minimum standards, and using the TOR to articulate very clearly the purpose of the evaluation.

Fundamental misunderstandings of recommendations and lessons learned prevail. Recommendations were often found to be disconnected from the preceding sections, drawing on the personal knowledge or opinions of the evaluator(s). Lessons learned generally perform even more

poorly overall. Indeed, the prevalence of misunderstanding of the lessons learned element of reports might suggest that it is a central candidate for explicit efforts to raise the awareness of both evaluation managers and evaluation teams about what lessons learned are.

The qualitative approach is a viable and useful way forward for the UNICEF Global Evaluation Report Oversight System. The experience of this meta-evaluation has found the qualitative approach to not only have been capable of generating analysis equivalent to or richer than the previous quantified methodology, but that it also enabled reviewers to provide more useful, accurate and constructive feedback to UNICEF managers in a range of contexts.

There is work to be done in supporting evaluation managers to mainstream human rights, gender and equity. There are some profound challenges in the interaction between HRBAP and robust evaluation that are unlikely to be addressed through simply revising weaknesses in the review tool.

Lessons Learned

These three lessons learned have been generated through the analysis of the core-evaluation team and adapted based on responses from UNICEF Evaluation Office and regional offices.

There is great value to be gained from blending the skills of evaluation teams. Complex rights-orientated evaluations have been delivered successfully where both evaluation skills and sector-knowledge has been present on the team. With an apparent shortfall in both evaluators with rights-knowledge and rights-specialists with evaluation skills, it would seem that value could be delivered from creating evaluation teams that mix these two skillsets as an alternative to the more traditional international/local knowledge blend.

Developing strong international frameworks provides a platform for stronger, more rights-orientated and more useful evaluation. Frameworks such as Child Friendly Schools and Core Commitments to Children empower evaluators to better manage rights issues within their evaluations. Stronger, more useful evaluations contribute to enhancing knowledge across these sectors and thus strengthening the frameworks themselves: thereby contributing to the creation of a 'virtuous spiral'.

Co-designing the methodology and investing in the development stage of evaluation-quality-reviews delivers a strong return in performance. Misunderstandings that could have become a problem at the analysis stage were eradicated early on through face-to-face working between UNICEF and IOD PARC teams. This also had the benefit of attenuating different interpretations of ratings by reviewers. Reviewers themselves had a chance to test and to help refine the review tool before it was finalised and deployed. This had the benefit of working through many possible scenarios and ultimately contributing to a more universally usable tool.

Recommendations for the UNICEF Evaluation Office

Recommendations have been generated through the analysis of the core-evaluation team and adapted based on responses from UNICEF Evaluation Office and regional offices.

Focus on delivering a strategy for more consistently high-quality terms of reference within the decentralised evaluation function. The significance of terms of reference in the current performance of evaluation reports has been a consistent presence in all stages of this meta-evaluation. This strongly suggests that it is a priority area for action that could deliver wide-ranging gains.

Continue development of the qualitative approach to deliver the Global Evaluation Report Oversight System. The findings of this report would appear to suggest that the flexibility of the review tool makes it more easily adaptable in a relevant way to different contexts and levels of

evaluation. Further work to formalise the approaches developed here is thus likely to be of benefit more widely within the organisation.

Develop and communicate an integrated HRBAP mainstreaming strategy for the evaluation function. The review tool used for this evaluation needs to be enhanced in terms of human rights, gender and equity assessment. However, this needs to be part of a more systemic strengthening of HRBAP within the evaluation function. A strategy and the tools for doing so are likely to have the most profound benefit if they are communicated effectively to evaluation managers and integrated as a central part of terms of reference for future evaluations.

Invest in clarifying results frameworks for challenging thematic areas, work to evaluate these at higher levels, and contribute this knowledge to international attempts to develop coherent frameworks. Not all cutting-edge themes and sectors benefit from a legacy of clear logic models or outcome and impact level indicators. Where this is the case, there appears to be a tendency to focus instead on output level evaluations that consistently deliver little in the way of real value. Thus, it might be a better approach to focus instead on doing a lesser number of higher-level evaluations in challenging areas and using these as a platform to invest in clarifying the results framework

Revisit how the requirements for recommendations and lessons learned sections are communicated to evaluators and managers, particularly in relation to policy-level evaluation. The Evaluation Office has developed an increasing array of communications tools to reach out to the decentralised evaluation function (such as MyM&E). It would appear to be timely to consider how these can target the capacity of policy evaluation managers and evaluators to deliver better recommendations and correctly-identified lessons learned.

Recommendations for Regional Offices and Country Offices

Recommendations for regional and country offices have been combined because they largely require joint working in order to deliver successful evaluation reports.

Cooperate on delivering basic quality assurance at the TOR stage and draft report stage of evaluations. Systematically applying even the most basic quality assurance checks to make sure that terms of reference and evaluation reports contain the essential elements can already make a substantial difference to the quality of reports. Regional and country offices should explore options for delivering these basic checks through systematically reviewing evaluation TORS and draft reports at the regional level.

Attempt to focus on fewer and better evaluations that deliver strategic priorities. The evidence presented here suggests that where it can be done then there may be significant gains in terms of evaluation report quality. The aid effectiveness principles of harmonisation, alignment, managing for results and national ownership can all be invoked to get donor support for using the IMEP to prioritise and focus on strategically important evaluations. Transitioning to strategic evaluation will also open up opportunities to engage in more joint and country-led evaluations.

Create cross-UNICEF pollination around upstream evaluations, and explore the options for multi-country approaches. Within regions, delivering at least one example of a successful and relevant multi-country evaluation is likely to enhance the wider evaluation function in these countries at the same time as influencing policy and demonstrating the value-added of the organisation. If this can be linked to the international agenda then it also provides a strong platform for the cross-pollination of knowledge within UNICEF.

Deliver extra support to country-led evaluations. Aid-effectiveness principles of the Paris Declaration and Accra Agenda for Action are increasing the prevalence of county-led evaluations. Concomitantly, results based management is a major entry point for strategic engagement between

UNICEF and national governments. Thus, it is recommended that Country and Regional Offices consider investigating the challenges experienced by country-led evaluations within their contexts and responding accordingly.

Potential areas of focus for improving evaluation report quality in specific regions

Based on our reading of the evidence, the following suggestions are offered as to possible focus areas for individual regions to move forward on enhancing the quality of evaluation reports.

TACRO	Focus on enhancing the processes for developing and communicating recommendations that are used by evaluators
CEE/CIS	Focus on maximising experience with rights issues by increasing the involvement of evaluators and tackling higher-level results
EAPRO	Focus on more consistently structuring evaluation reports and eradicating unsatisfactory conclusions sections
ESARO	Focus on enhancing the essential ingredients of a good evaluation through improving the quality of terms of reference
MENA	Invest in developing access to a wider cohort of evaluators with relevant contextual knowledge and experience
ROSA	Focus on quality assuring the analytical capability of evaluators by requiring detailed methodologies to be provided in inception reports (including analytical frameworks and processes)
WCARO	Explore options for sharing and learning from regional examples that are outstanding, particularly in relation to mixed methods
Corporate	Focus on sharing lessons for consistent evaluation with regions and try to maintain this consistency with a higher rate of knowledge generation

Contents

Figures

Tables

Introduction

UNICEF Evaluation Office (EO) has put in place a Global Evaluation Report Oversight System to monitor the impact of efforts to strengthen the UNICEF evaluation function globally. The system consists of rating evaluation reports commissioned by UNICEF Country Offices (CO), Regional Offices and HQ divisions against the UNEG/UNICEF Evaluation Report Standards.

Reports are made available in the UNICEF Global Evaluation Database (GED).This was created on the UNICEF Intranet in 2001. In June 2002, it was made available to the public on the UNICEF Internet website. It currently contains around 3300 records.

Background and purpose of the Global Evaluation Report Oversight System

UNICEF holds a long-standing commitment to independent assessment of the quality of evaluation reports produced by its country and regional offices all over the world, as well as HQ divisions. Most notably, this has included large-scale meta-evaluations, including Victora *et al* (1995), Watson *et al* (2004) and Noij *et al* (2010).

The quality review that is the subject of this report sits within this tradition. A central purpose of reports from previous years (and the processes that generated them) was to provide an accountability mechanism for the evaluation function, and to inform efforts to strengthen evaluation in line with the UNICEF Evaluation Policy. This report further extends the reach of this purpose.

The main purpose of this quality review process is to provide decision makers in UNICEF with information about evaluation reports that better supports using and improving the knowledge generated by the evaluation function. Whereas previous processes sought to create a 'reference point' for understanding the state of evaluation; this quality review seeks to actively inform the enhancement of evaluation report quality. Furthermore, it seeks to go beyond raising awareness of quality issues, and demonstrate the implications of trends in evaluation quality on the usability of knowledge in the pursuit of delivering results for children and women.

Scope of the review

This quality review process covered all evaluation reports submitted to the UNICEF Global Evaluation Report Oversight System for 2009. Reviews, research and other types of reports were excluded, as were evaluation reports undertaken by offices or divisions that were not submitted to the UNICEF Evaluation Office.

The quality review tool assesses the evaluation report as a *standalone document*. No additional investigation is made into the use, conduct or context of an evaluation. The standards against which evaluation reports are assessed are set by the UNICEF deployment of the United Nations Evaluation Group (UNEG) global evaluation report standards.

Objectives of the review

The overall objective is to assess and rate the quality of evaluation reports commissioned by UNICEF in 2009 using the UNEG/UNICEF Evaluation Report Standards.

Specific objectives

- Review and rate (with justifications) the quality of the main elements of evaluation reports, including structure, context, purpose, methodology, findings, conclusions, recommendation and lessons learned;
- To provide constructive feedback for evaluation commissioners to improve future evaluations;

- To provide a global analysis of key trends, strengths, weaknesses, and lessons of UNICEF evaluation reports; and
- To provide actionable conclusions and recommendations to improve the quality oversight system and systemic quality of the evaluation function.

Review Methodology

The meta-evaluation considered all of the screened reports that had been submitted as evaluations to the UNICEF Global Evaluation Report Oversight System for 2009. These reports were subjected to an initial filtering process by an evaluation expert to remove from the sample frame those reports that could easily be considered as being a form of assessment other than evaluation according to UNICEF standards.

> *"An evaluation is defined as an assessment, as systematic and impartial as possible, of an activity, project, programme, strategy, policy, topic, theme, sector, operational area and institutional performance. It focuses on expected and achieved accomplishments, examining the results chain, processes, contextual factors an causality, in order to understand achievements or the lack thereof. It aims at determining the relevance, impact, effectiveness, efficiency and sustainability of interventions."* (UNICEF, 2010 p5)

Reports were further filtered according to this definition by each reviewer prior to review. This process resulted in 99 full reviews. Of these, three were subsequently reclassified by the UNICEF Evaluation Office as being assessments other than evaluations. Thus, this meta-evaluation report draws on the complete sample frame of the remaining 96 reviews.

Each review was undertaken by an evaluation expert familiar with previous meta-evaluations of UNICEF evaluation report quality. All reviewers participated in a co-design workshop that enabled a common understanding of the standards to be reached based upon practical application to example reports and dialectic plenary discussion.

Following the completion of each review format, three further levels of assurance were applied. The first was a basic 'completeness' check to ensure that all relevant information had been provided. Secondly, a peer review by a member of the core evaluation team was undertaken on selected reports for each reviewer, with any disparities between results flagged to the original reviewer for clarification and/or amendment. Finally, the UNICEF Evaluation Office were able to exercise the right to challenge any review on technical grounds, with comments provided to the original reviewer for consideration.

UNICEF Evaluation Report Standards

In line with its international commitments, UNICEF has adopted a contextualised version of the United Nations Evaluation Group evaluation report standards. These identify 42 standards arranged under 8 main sections: report structure, object of the evaluation, evaluation purpose objectives and scope, evaluation methodology, findings, conclusions and lessons learned, recommendations, and gender and human rights.

These standards inform the UNICEF Global Evaluation Report Oversight System. One of the four main purposes of the Global Evaluation Report Oversight System (GEROS) is to assess the quality of evaluation reports: the component under which this meta-evaluation sits. GEROS establishes the standard against which this meta-evaluation assesses evaluation reports to be satisfactory or not.

> *"An evaluation report is assessed as satisfactory when it is a credible report that addresses the evaluation purpose and objectives based on evidence, and therefore can be used with confidence."* (UNICEF, 2010 p3, emphasis original)

In addition, the GEROS provides the basis for the development of a review tool by establishing five core 'elements' of an evaluation report that are required to respond to the eight adapted UNEG standards:

- Well structured, logical and clear
- Clear and full description of the 'object' of the evaluation
- The evaluation's purpose, objectives and scope are fully explained
- Appropriate and sound methodology
- Findings, conclusions and recommendations are based on evidence and sound analysis

Review tool, process and analysis

The full review tool is presented in the Annexes. This was co-designed by UNICEF and IOD PARC based upon the UNEG/UNICEF standards, lessons from previous global and regional meta-evaluations, and the Global Evaluation Report Oversight System framework. This review tool primarily adopts a qualitative approach to rating evaluation reports against the overall standard of confidence. It pursues a systematic process of aggregating qualitative ratings of 58 guiding questions about different aspects of an evaluation report into six sections, and then into a final overall assessment.

The six sections of analysis are based on the five core elements identified in the Evaluation Quality Assurance System, with the further separation of recommendations and lessons learned as a separate lens of enquiry. The six sections of the review tool are:

- Object of the evaluation
- Purpose, objectives and scope
- Evaluation methodology, gender, human rights and equity
- Findings and conclusions
- Recommendations and lessons learned
- Report is well structured, logical and clear
- (Plus additional information and an overall reaction)

This qualitative approach is designed to enable reviewers to provide useful analysis across the range of evaluation contexts encountered; and constructive feedback to improve future evaluation reports. Each question, each section and the overall report are given a rating of either 'very confident', 'confident', 'almost confident' and 'no confidence' (where relevant, a N/A option is also provided).

In addition to ratings, commentary is provided against each rating, suggestions for future improvement provided for each section, and executive feedback provided for each section and the overall report. The complete review process generates three types of data: a report typology, a series of ratings, and a structured set of discussion text.

Report typology

Evaluation reports are initially classified according to commissioning agent and the UNICEF evaluation typology. This allows analysis according to various evaluation report characteristics. The seven typology criteria are: geographical coverage, management, purpose, result level, MTSP[1] correspondence, level of independence, and timing.

Assessment ratings

Assessment ratings were given against 58 guiding questions, six sections, two meta-questions, and an overall view of the report. Each 'aspect' was given one of four qualitative ratings: two ratings accorded to the reviewer finding that a UNICEF manager could confidently act on that aspect of the evaluation

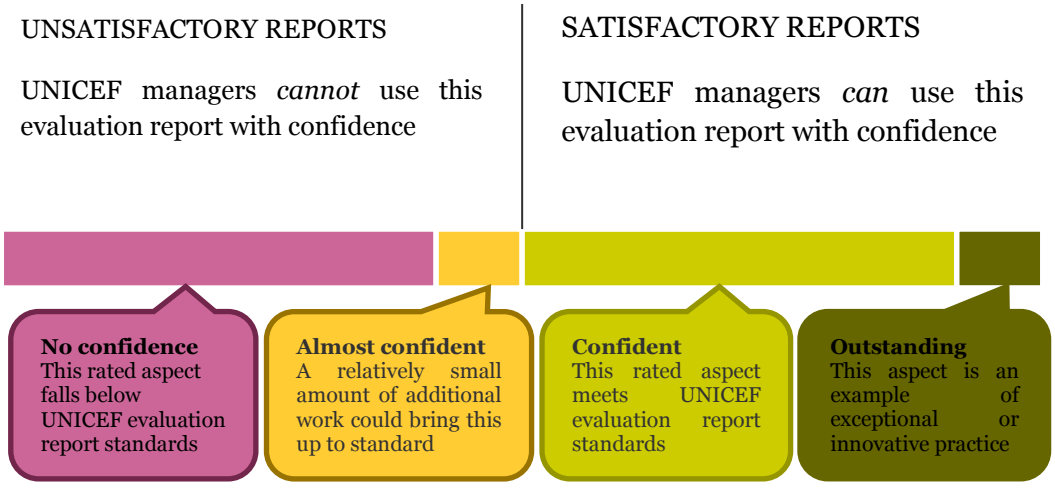
[1] Medium Term Strategic Plan

report, and two ratings accorded to the finding that a UNICEF manager could not be confident in that aspect of the evaluation report.

Each rating was informed by three factors: a prompting question in the review tool; a 'confidence-to-act' test, and any ratings in the level of analysis below. In other words, the overall rating was informed by the six section ratings, and each section rating was informed by the question ratings clustered under that section. Within this 'aggregating' system, each individual rating had the following properties:

1. **Ratings are absolute measures.** *Each report rating is made against the UNICEF Evaluation Report Standards. It is therefore possible for all reviewed reports to achieve an 'outstanding' rating, or all reports to be classified with a 'no confidence' rating. There is no relative assessment within a 'cohort' of reports.*
2. **Ratings are not compoundable.** *Ratings are not 'quantitated' representations of qualitative assessments and cannot be used for calculating 'mean' ratings. Thus, a particular report may have a mix of 'outstanding' and 'confident' section ratings, but these do not 'automatically' add up to give a particular overall rating. Rather, the different levels of ratings (questions, sections, and overall) allows the reviewer to build up a holistic overall view on the quality of an evaluation report. Similarly, overall report ratings are not compoundable: thus, a particular region cannot be given an 'average' score.*
3. **Ratings are not distributed on an even scale.** *There are four possible ratings available to each reviewer. The main distinction between these is between 'confidence-to-act' and 'no confidence'. An additional two ratings – 'outstanding' and 'almost confident' – allow a reviewer to provide a more nuanced response to highlight a narrow selection of reports (or elements of reports) that can be flagged as either best practice examples, or requiring just minor amendments in order to reach a satisfactory standard.*

Figure 1: visual representation of ratings used



Justification, feedback and recommendations

The second main dataset to be generated is qualitative text linked to each rating. The primary purpose of this text is to enable evaluation managers to understand a) why their reports received particular ratings, and b) how to improve future reports. Executive summaries for each section and each report also provide rapid feedback to UNICEF senior managers that highlights the strengths, weaknesses and lessons for each report. Qualitative analysis of this rich dataset is used to inform the trend analysis.

Meta-evaluation Methodology

Trend analysis

The review process generated an extensive dataset to inform the trend analysis, consisting of 1,152 individual pieces of evaluation typology data, 6,432 individual ratings, and 6,912 sections of qualitative text (approximately 140,000 words). In order to distil the key findings from this data, the following multi-stage process was adopted.

1. All reviewers submitted analysis identifying the key themes (strengths, weaknesses and issues) that they had encountered in the evaluation reports that they reviewed, and in the review tool itself;
2. The core evaluation team synthesised reviewer feedback into the core issues identified under each section of the review tool;
3. All reviews were completed within an Excel template: these were then able to be aggregated into a single workbook using the DigDB plug-in for Excel[2];
4. A master worksheet was developed that pulled rating data and selected feedback text from the reviews into single spreadsheet for analysis;
5. Rating data was disaggregated using evaluation typology data to produce a series of distribution bar graphs that visualised the 'spread' of ratings for different aspects of the evaluation reports (examples of these can be found in the findings section);
6. Qualitative text comments were disaggregated according to overall report and section ratings and combined into blocks of text for analysis;
7. Blocks of text were subjected to inductive coding[3] by using the Word 'find and replace' tool to remove frequently-occurring words (such as "evaluation", "report" and "satisfactory") and the Wordle[4] common English words filter (such as "the", "and" and "a").
8. The inductive coding generated a list of most-frequently occurring words and issues disaggregated according to the ratings given. This was compared to the synthesis of reviewer comments in order to identify the prevalence of the main trends.
9. A selection of reports with representative or interesting ratings (or those flagged by reviewers) were then revisited in or to triangulate the findings of the coding process, or to add substantiating details to these findings.

Changes from previous years

Readers of previous UNICEF meta-evaluations will have noticed that the methodology adopted here is a substantive departure from previous approaches. Previously, 22 evaluation standards were allocated 1-5 scores that could be combined to provide an overall 'number' for an evaluation report. This approach used a fixed weighting system to generate a report's final score and did not allow for contextual adaptation by reviewers. The core trend analysis primarily focused on quantitative processing of these scores to reveal common themes.

Feedback from UNICEF indicated that evaluation commissioners also had a tendency to focus on 'scores' rather than the underlying issues identified by reviewers. It is intended that the current methodology should contribute to a richer set of learning by increasing the focus on the reasons *why* reports are rated as confident or not; and deliberately avoiding the impression that the difference between satisfactory or unsatisfactory reports is a matter of scoring a few more 'points' here or there.

Limitations of the approach

As with all evaluations, the approach taken by this meta-evaluation is subject to real-world constraints, including resources and time. This requires a balance between what is ideal and what is

[2] DigDB automatically runs a 'combine' macro, a free example of which can be found at http://excel.tips.net/Pages/T002409_Merging_Many_Workbooks.html
[3] For an explanation of inductive coding see http://iodparc.com/resource/qualitative_indicators.html
[4] www.wordle.net

possible: and creates a number of limitations that must be taken into account. Where relevant, these limitations have been mitigated through triangulation of both the review and trend analysis processes in order to constrain opportunities for inconsistency or information-loss.

The review process itself is subject to the limitation of only having access to the written evaluation report (and these did not always include the Terms of Reference). As a direct consequence of this, the findings and conclusions drawn can only be applied to an evaluation report, and not to the evaluation itself. Furthermore, it should be recalled that the reports being assessed were published in 2009, whereas the final Evaluation Report Standards that inform the review tool were finalised in 2010. So to some extent this meta-evaluation is assessing yesterday's reports against today's standards. However, whilst some details have changed, the core values that underpin the new evaluation report standards *are* consistent with previous iterations of those standards, largely removing any case for the possibility of under-rating.

Qualitative analysis (as with all analysis) requires for judgements to be made in identifying the important indicators and trends contained within the dataset generated by the review process. An ideal approach may have been to separately consider each of the 14,000 lines of text written in order to cluster points around the main trends: unfeasible in this instance. Thus, the approach adopted is limited to being able to identify only the 'headline' findings, with the possibility that more nuanced or infrequently occurring issues exist for individual readers to find within the reviews themselves.

Findings

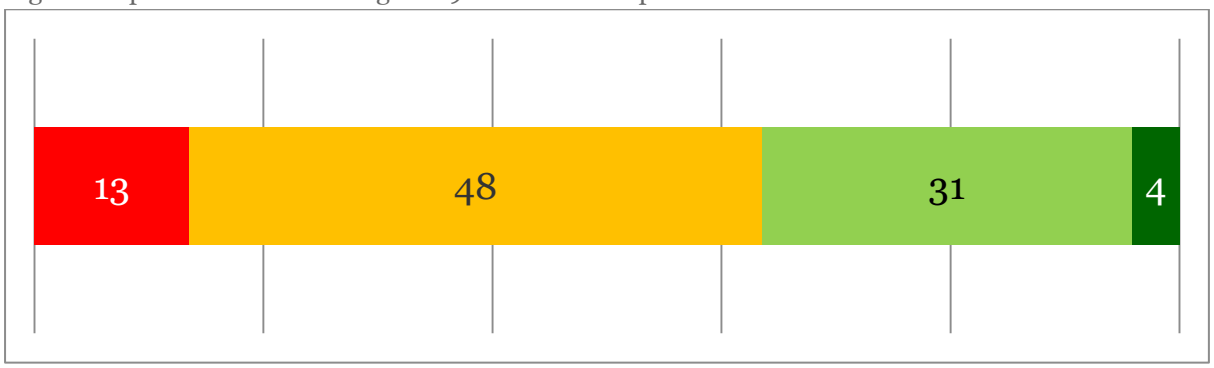
This report draws on 96 reviews of evaluation reports submitted from the following regions:

The Americas and Caribbean	12	Middle East and North Africa	6
Central and Eastern Europe, Commonwealth of Independent States	16	South Asia	11
East Asia and the Pacific	17	West and Central Africa	12
East and Southern Africa	19	Headquarters Divisions	3

Overall findings on quality

The meta-evaluation found that 36% of reviewed evaluation reports met the UNICEF standards to a degree that could be considered satisfactory. Whilst the remaining 64% of reports were rated as unsatisfactory, the vast majority of these (exactly half of all reports) could have been improved to a satisfactory level with just a little more work. A regional breakdown of ratings is provided in Annex 8.

Figure 2: spread of overall ratings for 96 evaluation reports



Thirteen reports were deemed to be to a level where there could be no confidence to act, with the majority of these being unsatisfactory in all sections of the assessment. Overall, four evaluation reports were flagged as outstanding best practice, although six more achieved a very confident rating in one or more of the review sections. The outstanding evaluations were Thailand's Evaluation of Children and the 2004 Indian Ocean Tsunami, Guinea Conakry/Guinea Bissau's Evaluation of WASH Activities, Timor Leste's Evaluation of the UNICEF Education Programme, and Uzbekistan's Evaluation of the Family Education Project.

Qualitative analysis reveals that those reports rated as *outstanding* were noted primarily for developing clear and thorough conclusions. These were grounded in robust evidence that was transparently generated through credible analysis, and framed by a clear set of objectives.

Reports rated as *confident* were particularly noted for clear and well-presented findings. These were still credible, but did excel to the same degree as outstanding reports in the use of evidence, or the development of robust conclusions and recommendations. Satisfactory reports were noted for tending to benefit from a strong set of evaluation questions.

The half of reports that were rated overall as *almost confident* systematically fell short of the standards in relation to weakness with findings, conclusions and recommendations. Despite having reasonably clear purposes and appropriate methodologies, the data lacked credible analysis and this undermined the evidence available to fully support latter sections of reports.

A lack of evidence contributed strongly to incoherent findings, conclusions and recommendations that were rated as *not satisfactory*. However, these also tended to have more profound issues in early

stages of the report, including unclear or unstated purposed and inadequate methodologies. Weak terms of reference were often suspected of contributing to these problems, although the infrequency with which these were attached to reports made it challenging to verify this in individual cases.

Credibility and coherence of report sections

Whilst determining the overall rating of a report in the final stage of each review process, reviewers were prompted to consider how credible the sections of the evaluation report were separately, and how well they held together as a whole. This rating data suggests that the final overall score tended to be more closely aligned to the quality of individual sections. The majority of reports appear to be more satisfactory in terms of coherence than their overall ratings, suggesting that individual sections within the report are causing problems.

This pattern is different, however, for the weakest evaluation reports: where lack of structure and coherence appeared to cause very real problems. These reports relied upon individual elements within the document to add value, rather than contributing as a whole. Qualitative data on explanations for these weaknesses that could be found evident within the reports was also captured as part of the review process. Analysis of this data highlights how flaws early-on in the evaluation process erodes the performance of the evaluation report as a whole.

Reviewers noted in particular that unclear objectives were a major contributor to poor report quality, as was having an unclear purpose, inadequate evaluation questions, or missing evaluation criteria. In nearly all of these cases it was found that either poor quality terms of reference contributed to the weaknesses evident in evaluation reports, or (where TOR were not attached) the issues were ones that should have been specifically clarified in the TOR for that evaluation. Further analysis of this issue is included under the findings for Section F, below.

Figure 3: word cloud of qualitative feedback on the causes of weak evaluation reports (larger words appear more frequently in the text)

Geography

The vast majority (95%) of evaluation reports covered national or sub-national areas, split more-or-less evenly between the two. One report was classified as being *regional*. This was rated as 'almost confident', although the small sample size means that no findings can be drawn from this. Four evaluation reports were classified as *global*, another small sample size. However, these four global reports stand out as being consistently of satisfactory quality: with only slight weaknesses noted in relation to the 'purpose' and 'methodology' sections.

The bulk of evaluations were found to be stronger in the early review sections (description, purpose, methodology) than the latter (findings, recommendations, structure). This was consistent across both

national and *sub-national* evaluations. However, taken as a whole, national evaluations tended to be viewed with greater confidence than sub-national evaluations, with almost twice the number of reports be rated as satisfactory.

Management

The majority of reports (57%) were classified as *UNICEF-managed*, around 17% of evaluations were *jointly managed with another development partner*, seven were *country-led* and one was *jointly managed with another UN agency*. The remaining 17 reports were unable to be classified from the report contents, with nearly 90% of these unclassifiable reports being rated as unsatisfactory.

The single UN jointly managed report was rated as satisfactory across all sections[5]. Whilst the seven country-led evaluations were rated poorly overall (less that 30% were considered satisfactory), they were noticeably stronger than other evaluation types in relation to 'description of the evaluation object'. However, very weak ratings for recommendations and structure sections, along with unconvincing performance in the other sections, appear to have eroded their overall performance.

Of the remaining evaluations, UNICEF-managed evaluations are rated as consistently more satisfactory than those managed jointly with another development partner (although still less than half of UNICEF managed reports are considered to be less than satisfactory). To some extent this may be considered to be symptomatic of the compromises necessary when multiple stakeholders are managing an evaluation process. Nevertheless, some fundamental elements of a satisfactory evaluation report – such as purpose and methodology – are rated as particular areas of concern with jointly managed evaluation reports.

Purpose

The review process allowed for multiple purposes to be allocated to each individual evaluation. A considerable majority of evaluations were classified as relating to *pilot* (19) or *at scale* (35) evaluation objects, with only seven *policy* level evaluations. Seven evaluations were classified as *humanitarian*, with two *real-time* evaluations: one of which was rated as satisfactory and one as unsatisfactory. In addition to this, 22 evaluations were classified as *project*, 19 as *programme* and four as *country programme* evaluations.

Among the reports reviewed, the four Country Programme Evaluations stand out as consistently strong across all sections of the review. Humanitarian evaluations also registered as largely satisfactory across all review sections. All other levels of evaluation range between more than 50% and more than 70% unsatisfactory. Around half of all at-scale, policy, project and programme reports suffered from unclear purpose or objectives. The weakest section was 'recommendations and lessons learned'; with only programme evaluations registering at least 50% of reports as satisfactory for 'methodology', 'structure' or 'findings and conclusions'.

Result

Approximately one quarter of reports sought to determine *outcome* level results, one third primarily considered *output* results, and the significant remainder were *impact* level evaluations. One report could not be classified. Of the classified reports, output level reports (30% of all reports) displayed concerning weakness across all review sections such that 90% were rated as unsatisfactory. Outcome and impact reports were consistent across both result levels, with around 50% being satisfactory overall, strong opening sections and weaker recommendations sections. Only impact reports included those that were considered to be outstanding best practice.

MTSP correspondence

The classification of alignment with Medium Term Strategic Plan areas[6] allowed for multiple responses. A total of 14 evaluations were classified as *multi-sector*, 23 as *cross-cutting*, and 5 as

[5] Azerbaijan UNDAF Evaluation

[6] The five MTSP areas are: 1/ Young child survival and development, 2/ Basic education and gender, 3/ HIV/AIDS, 4/ Child protection, and 5/ Policy and advocacy.

organisational performance evaluations. Thus, the remaining 56 can be reasonably considered to have focused on a single MTSP area.

Cross-referencing ratings with MTSP-correspondence reveals a consistent story across all sections of the review. Multi-sector and cross-cutting evaluations register strongly in all sections, with at least (or nearly) half of these reports being rated as satisfactory. Conversely, organisational performance evaluations were continuously ranked as unsatisfactory across all review sections (with the notable exception of the Global Evaluation of DevInfo).

From among the MTSP focus areas, *young child survival and development* suffered from particularly weak evaluation reports, less than 20% of which were considered to be satisfactory (albeit two being rated as outstanding[7]). *Policy and advocacy*, and *child protection* were fairly robust in terms of 'description of the object' and 'purpose' sections, but rated very poorly in all other sections. It was concerning to note that not a single policy evaluation report was rated satisfactory in relation to 'recommendations and lessons learned'.

HIV/AIDS evaluation reports were assessed to be poor in relation to all sections of the review, in particular with regard to methodological rigour. At the other end of the scale, *basic education and gender* evaluation reports were found to be consistently the strongest of all the MTSP focus areas with more than half of reports rated as satisfactory in all sections except for 'recommendations and lessons learned'.

Level of independence

The level of evaluation independence could not be ascertained for ten of the reports, and all-but-one of these were ranked as unsatisfactory. The majority of the remaining reports were classified as either *independent internal* or independent external, more-or-less evenly. Five reports were classified as self-evaluations, which, although performing fairly unsatisfactorily overall, were notable for out-performing other evaluations in relation to 'description of the object' and 'methodology' sections.

In all other reports, independent external evaluations tended to be rated as slightly more satisfactory than independent internal evaluations. This disparity became more pronounced in relation to 'recommendations' and 'report structure' sections of the review.

Stage/timing

The timing classification in UNICEF relates to whether an evaluation is *formative* or *summative*. This meta-evaluation encountered a fairly even number of both, with 42 formative evaluations, 52 summative evaluations and two unclassified reports. The ratings articulate a consistent story, with between 50% and 150% more summative reports than formative reports being classified as satisfactory in each and every section. Once again, the weakest sections for both stages of report related to 'recommendations' and 'report structure'.

Section A: Object of the evaluation

"The report describes the object of the evaluation including the results chain, meaning the 'theory of change' that underlies the programme being evaluated. This theory of change includes what the programme was meant to achieve and the pathway (chain of results) through which it was expected to achieve this.

The context of key social, political, economic, demographic, and institutional factors that have a direct bearing on the object is described. For example, the partner government's strategies and priorities, international, regional or country development goals, strategies and frameworks, the concerned agency's corporate goals and priorities, as appropriate."

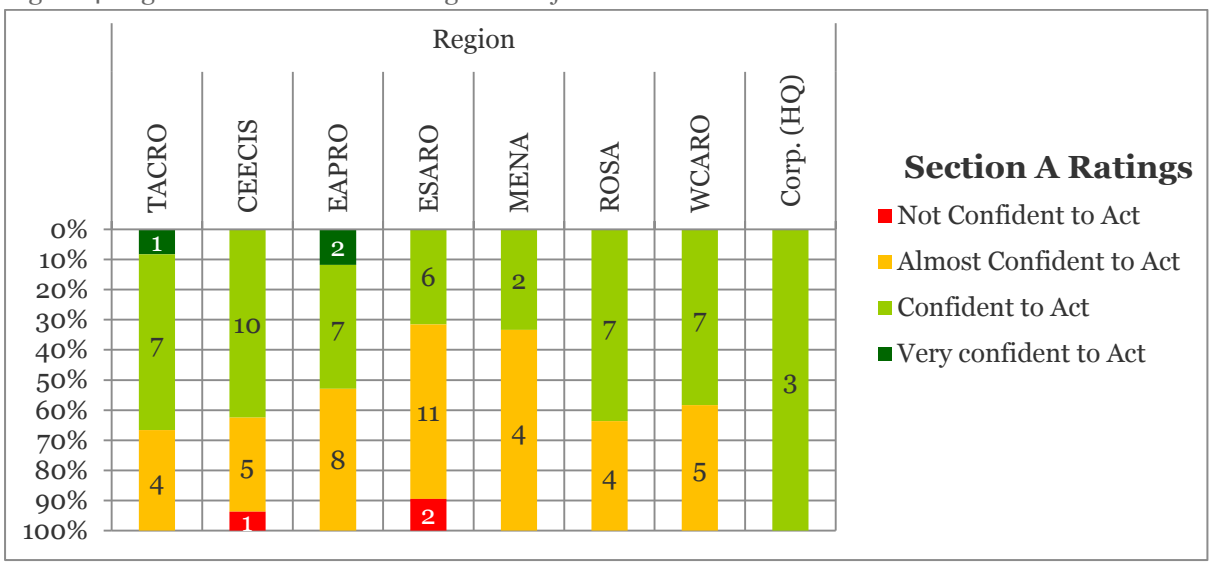
[7] Guinea-Conakry/Guinea-Bissau Evaluation of WASH activities, and Uzbekistan Evaluation of the Family Education Project.

The most frequent observation by reviewers is that evaluation reports fail to explicitly articulate the results chain of the evaluated object. Many reports were still able to present a satisfactory context through inclusion of other information, but the consequence of this issue is that a majority of reports do not appear to be guided by the logic of the programme or project being evaluated.

Another frequent observation was that there is a tendency to provide general information about a country or the implementation context of the evaluated object, rather than analysis that can shape the evaluation purpose, objectives, and findings. In a number of instances these two issues left the impression that the evaluators did not have a strong understanding of either the evaluated object or the country context.

Although Section A was rated overall as the most satisfactory of all the review sections, further analysis reveals that this varied considerably by region, by MTSP-correspondence, by level of result, and by timing/stage. Formative, output, HIV/AIDS and young child survival evaluations were all rated more weakly in relation to the description of the object than their contemporaries. Reports rated *outstanding* on Section A were Colombia's Evaluation of Educacion en el riesgo de Minas, Indonesia's EFA Mid-decade Assessment, and Timor Leste's Evaluation of the UNICEF Education Programme.

Figure 4: regional breakdown of ratings for 'object of the evaluation'[8]



Outstanding reports were noted as tending to link the description of implementation status to both the wider context and the purpose of the evaluation. These top-performing reports also clearly articulated or reconstructed the results chain of the evaluated object. This was accompanied by clear analysis of the different stakeholders, along with explicit descriptions of their roles and contributions. Evaluations of education-related projects and programmes appeared to rate particularly well on description of the context.

Reports that were rated as *confident* differentiated themselves by the clarity of the description present in the context section. These reports provided the reader with information about the results chain of the project, programme or policy being evaluated and, to a lesser extent, the contributions of different stakeholders

Almost confident reports tended to be unclear about, or to miss, the description of the underlying logic of an evaluated object. Although description of the context was included, intended results were not sufficiently described; and the reports lacked important information about stakeholders and implementation status.

[8] Regional breakdowns for other sections can be found in Annex 8.

The small number of reports that were rated as *not confident* on description of the object either provided no real contextual information, or only gave general country information that did not contribute any understanding of the results chain, stakeholders, or implementation status.

Section B: Evaluation purpose, objectives and scope

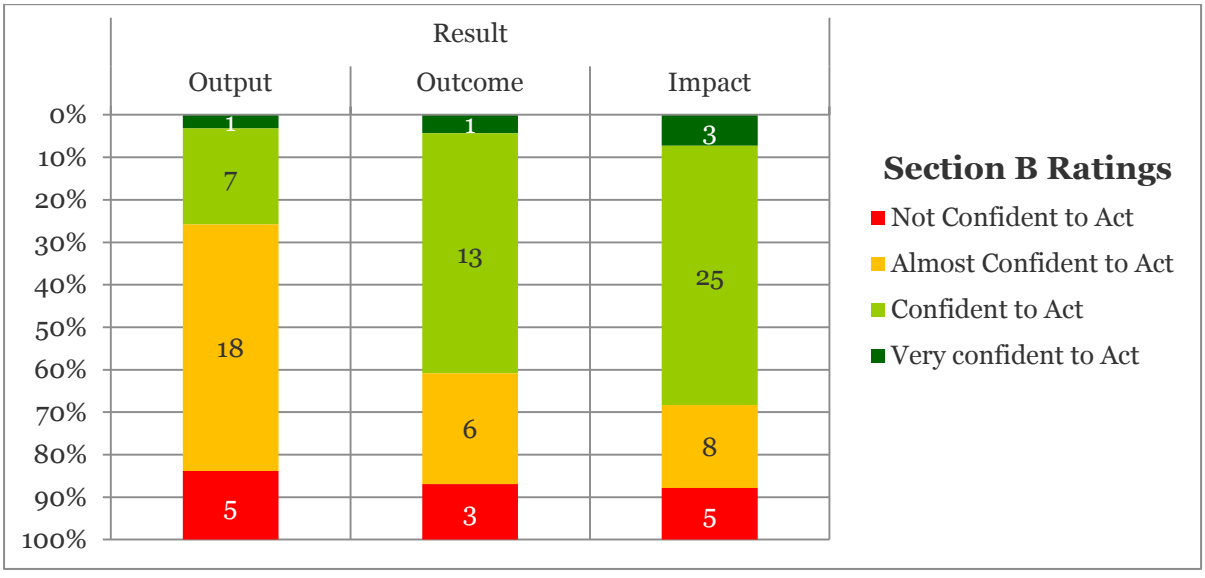
"The purpose of the evaluation is clearly defined, including why the evaluation was needed at that point in time, who needed the information, what information is needed, how the information will be used. The report provides a clear explanation of the evaluation objectives and scope including main evaluation questions and describes and justifies what the evaluation did and did not cover. The report describes and provides an explanation of the chosen evaluation criteria, performance standards, or other criteria used by the evaluators."

Reviews of the purpose, objectives and scope of evaluation reports revealed a number of underlying issues with the framing of evaluations. These appear to frequently manifest themselves through weak justification of evaluation criteria, and lack of consistent use of these criteria within evaluations. A common example of this is the citation of the OECD DAC criteria in the introduction, but the subsequent development of evaluation questions that correspond almost entirely to purely operation- or implementation-specific aspects of the evaluated object.

Once again, many reviews noted that the terms of reference play a central role in this trend. TORs that *were* present often have unclear or incorrect purposes (for example stating the purpose as to 'evaluate the programme in question'). TORs were also found to be setting evaluation questions that were inconsistent with the stated purpose, objective or scope. Another concern that was raised by some reports was that TORs appear to be framing purposes and objectives in a way that limits their learning value even if delivered outstandingly: evaluations are being used to check whether UNICEF is achieving planned targets rather than to identify problems and learn from them.

This latter issue perhaps also contributes to a final observation that there appears in many cases to be a heavy reliance on technical specialists, without blending specific evaluation skills into the resources committed to the assessment. This might go some way to explaining the significant disparity between output and formative evaluations, and outcome, impact and summative evaluations (that appear to be more likely to include evaluators). However, this cannot be the only explanation, as joint and country-led evaluations also rate poorly compared to their contemporaries: suggesting that the process of negotiation between stakeholders might dilute the clarity of purpose that is available to the evaluation team.

Figure 5: break down of purpose, objective and scope ratings by result level.



Despite these challenges, more than half of reports still rate as satisfactory (although there is still some regional disparity). The reports rated as being outstanding were Colombia's Evaluation of Educacion en el riesgode Minas, Guinea-Conakry/Guinea Bissau's Evaluation of WASH Activities, Timor Leste's Evaluation of the UNICEF Education Programme, and Uzbekistan's Evaluation of the Family Education Project.

Outstanding reports were noted primarily for having very strong evaluation frameworks. These clearly referenced the OECD DAC evaluation criteria in addition to identifying and integrating relevant rights instruments, such as the Core Commitments to Children. Reports rated as *confident* tended to have very clear purpose, objectives and scope. These guided evaluation criteria that mostly made explicit reference to the OECD DAC standards.

Evaluation reports rated as *almost confident* tended to still benefit from having usable purpose, objectives and scope. However, they did not translate these into a clear set of evaluation criteria that could then be used to guide the inquiry or analysis in the rest of the report. Not confident reports systematically failed to articulate clear purpose, objectives, scope, evaluation questions or evaluation criteria. Some of the possible explanations for these shortfalls have already been discussed above.

Section C: Evaluation methodology

"The report presents transparent description of the methodology applied to the evaluation that clearly explains how the evaluation was specifically designed to address the evaluation criteria, yield answers to the evaluation questions and achieve evaluation purposes.

The report presents a sufficiently detailed description of methodology in which methodological choices are made explicit and justified and in which limitations of methodology applied are included. The report gives the elements to assess the appropriateness of the methodology. Methods as such are not 'good' or 'bad', they are only so in relation to what one tries to get to know as part of an evaluation. Thus this standard assesses the suitability of the methods selected for the specifics of the evaluation concerned, assessing if the methodology is suitable to the subject matter and the information collected are sufficient to meet the evaluation objectives."

Ratings of evaluation methodologies were the first aspect of the review where just less than half of reports were considered to be satisfactory. Multiple reviewers found methodologies to be narrow and inadequately explained as a general rule, with many assumptions appearing to be made in the justification of particular approaches. This is manifested in terms of weak control of bias: with some reports doing no more than the evaluator providing personal reflections on a narrow set of interviews.

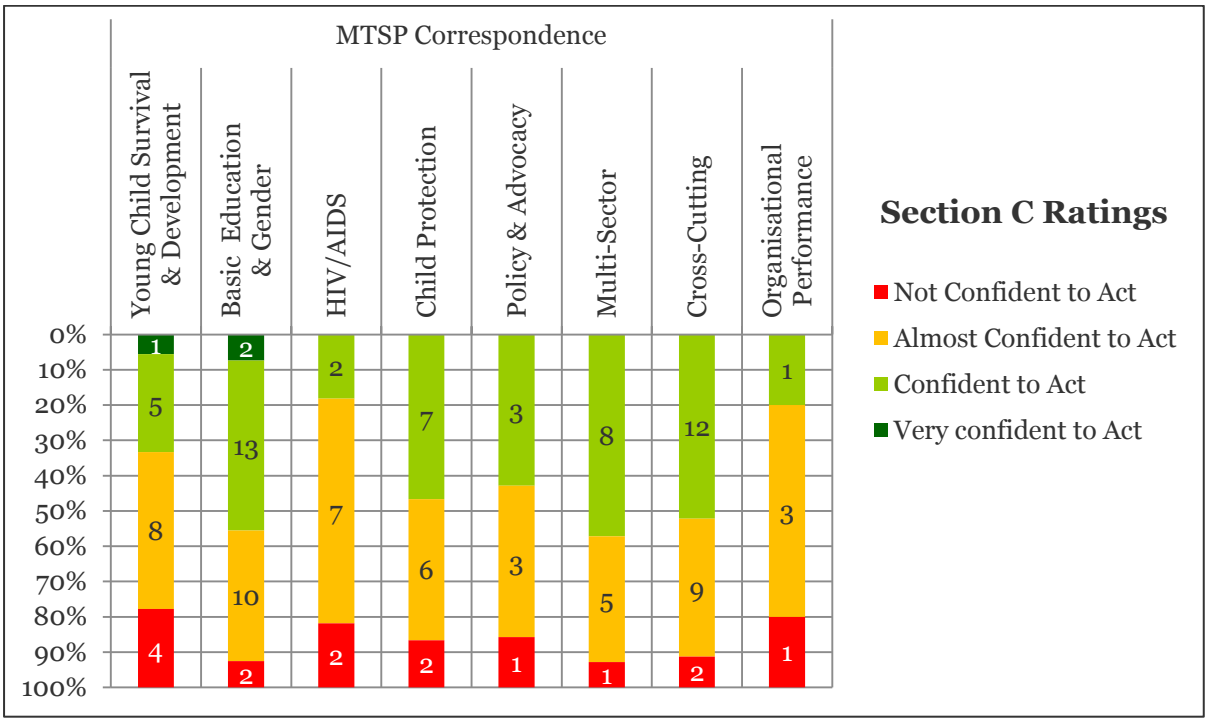
Overall, ethics remains a weak area in evaluation reports (a consistent observation in previous meta-evaluations). Most reports do not include any discussion on ethics, although it is sometimes evident from the approaches adopted by evaluators that ethical considerations had been borne in mind at some point. This once more raises the question of whether terms of reference need to specifically build-in a requirement for ethics and other weak areas, such as establishing counterfactuals and demanding cost-benefit analysis. Despite this overall weakness, however, three reports were noted for being outstanding in their dealings with ethics: Colombia's Evaluation of Educacion en el riesgo Minas, Dominican Republic's Evaluation of the Estrategia de Comunicación y Movilización Social, and Uzbekistan's Evaluation of the Family Education Project. This demonstrates that best practice examples *are* available for others to learn from within the UNICEF context.

A notable observation that a number of reviewers made was that the majority of reports appear to suggest that M&E systems for the evaluated object are either weak or virtually non-existent. The logical frameworks, where they are included, can often appear to be confused, with poorly formed indicators and no baselines. This can create an immediate barrier to being able to deliver a strong evaluation (although it should also be recalled that the evaluation report standards do allow for an evaluation team having to manage around such issues).

Figure 6: ethics remains an area of concern in this meta-evaluation

A final observation picked up from reviewers and confirmed by the data is that there is a disparity between the performance of education-related evaluations and rights-orientated evaluations, particularly HIV/AIDS. This appears to be consistent across the review sections that rely on evaluation skills, such as methodology, lending credence to the previously discussed reliance of rights-orientated evaluations on technical sector specialists rather than evaluators. If this is the case, then it may be explained as much by a lack of rights knowledge among evaluators as by a lack of evaluation skills among sector specialists. Either way, it appears to highlight a potential need for enhancing the use of mixed teams.

Figure 7: breakdown of methodology ratings by MTSP-correspondence



The evaluation reports rated as *outstanding* on methodology differentiated themselves in the very clear articulation of limitations of the approach, data sources and links to the evaluation purpose and objectives. All of these required attention to a robust and appropriate underlying methodology. Reports rated *confident* were found to be clear and transparent, with consistent use of evaluation

criteria to develop appropriate methodologies that made some reference to limitations and ethical issues.

Those reports rated as *almost confident* tended to have a reasonable attempt to express a methodology, but this was either overly narrow or lacked important discussion on limitations or data sources. None of these reports addressed ethics issues. Finally, evaluation reports rated as *not confident* lacked description of methodology and data sources, were subject to overwhelming (and unexplored) limitations, or generally made inappropriate methodological choices that were delinked from the report objectives.

Section D: Findings and conclusions

"Findings respond directly to the evaluation criteria and questions detailed in the scope and objectives section of the report. They are based on evidence derived from data collection and analysis methods described in the methodology section of the report.

Conclusions present reasonable judgments based on findings and substantiated by evidence, providing insights pertinent to the object and purpose of the evaluation."

Reviews of the findings and conclusions elements of evaluation reports appear to tell two very different stories. Some evaluations were praised for collecting diverse datasets and large potential bodies of evidence. The best of these were able to convert this data systematically into evidence, findings and conclusions using strong and transparent analysis. Reports rated as *outstanding* in this regard were Guinea-Conakry/Guinea Bissau's Evaluation of WASH Activities, Indonesia's Mid-decade Assessment of Education for All, Nepal's Joint Evaluation of Education for All, and Timor Leste's Evaluation of the UNICEF Education Programme.

The majority of reports were unable to demonstrate this systematic use of evidence to construct robust findings and conclusions. Indeed, there appears to be a persistent problem in regard to data analysis, with reports not using data to its full potential or presenting processed information. Many reports were found to have presented the raw data in the findings section: eroding readability, creating gaps in the logical progression of the report and, misrepresenting the purpose of evaluation.

Where data was analysed, it was frequently not done so in a way that could answer the evaluation questions set at the beginning of the report. This was a trend that was continued in the discussion of conclusions: a significant number of reports did little more than present a summary of the findings that did not add value or deeper insight.

The importance of the 'evidence issue' is revealed in the qualitative analysis of feedback data. Reports rated as *outstanding* were noted for consistently generating clear evidence bases through systematic analysis of data, with logically derived findings and conclusions. These were linked back to the objectives of the evaluation through the consistent use of a strong evaluation framework throughout the reports. Those reports rated *confident* demonstrated the same characteristics, albeit to a lesser extent.

Unsatisfactory reports rated as *almost confident* still presented findings and conclusions, but failed to substantiate these through analysis of data or clear use of evidence. The weakest reports, rated *not confident*, lacked the use of evidence altogether. These tended to tell a story about the evaluated object, or rely purely on a narrow set of interviews, before offering personal opinions and subjective interpretation as a view of performance.

Section E: Recommendations and lessons learned

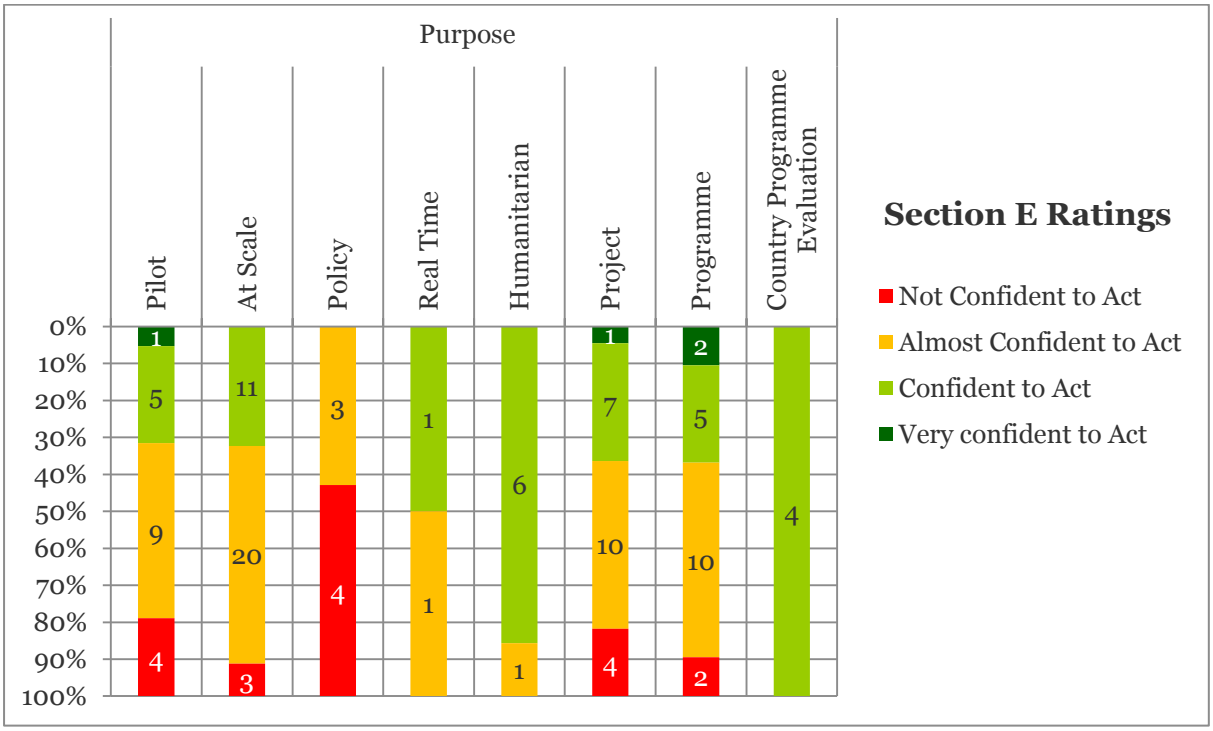
"Recommendations are relevant to the object and purpose of the evaluation, are supported by evidence and conclusions, and were developed with involvement of relevant stakeholders. Recommendations clearly identify the target group for each recommendation, are clearly stated with priorities for action, are actionable and reflect an understanding of the commissioning organization and potential constraints to follow up"

Recommendations and lessons learned were split from findings and conclusions as an assessment area. At the time of the review design there was some concern over the appropriateness of including recommendations in all evaluation reports[9]. Thus, reports that did not include recommendations or lessons learned were not rated for this section (and so absence of these elements did not reduce the overall rating of the report quality). Even so, recommendations and lessons learned was the weakest rated section in the review, with only one third of the rated reports being assessed as satisfactory.

Reviewers noted that it was often hard to see the link between recommendations and the preceding findings and conclusions. In some instances this appeared to continue the trend noted in the previous section of evaluators leaning primarily on their subjective opinion. In other instances, reports that were otherwise strong rated weakly where evaluators that were knowledgeable about the project introduced recommendations for which there was a weak evidence-base in the report.

Lessons learned proved to be even more problematic than recommendations. When they were found in reports, lessons learned were more-often-than-not found to be project or programme-specific observations and not generally applicable to other contexts. Perhaps the most concerning observation, however, was that all of the seven policy evaluations included in the review were rated as unsatisfactory in relation to recommendations and lessons learned. This is an issue of some significance in relation to UNICEF's commitment to more upstream working.

Figure 8: breakdown of recommendations and lessons learned ratings by purpose



[9] This has since been revised, as UNEG standard 3.16 states that both recommendations and lessons learned are an integral part of evaluation reports.

The three reports that were rated as *outstanding* appear to have been differentiated largely by the clarity with which well-prioritised and appropriate recommendations were stated. These reports were Colombia's Evaluation of Educación en el riesgo de Minas, Guinea-Conakry/Guinea-Bissau's Evaluation of WASH Activities and Timor Leste's Evaluation of the UNICEF Education Programme.

Some reassurance should also be taken from the findings that all Country Programme Evaluations and nearly all humanitarian evaluations were rated as *confident*. Reports rated as confident tended to have specific, relevant, realistic and actionable recommendations that were clearly ground in the preceding findings and conclusions. Conversely, *almost confident* reports lacked prioritisation or were based on poor evidence, findings and conclusions. *Not confident* rated reports had recommendations that were unclear and lessons learned that were incorrectly identified. These were often based on personal views, made weak use of evidence and were not prioritised.

Section F: Report structure, logic and clarity

"The report is logically structured with clarity and coherence (e.g. background and objectives are presented before findings, and findings are presented before conclusions and recommendations). It reads well and is focused."

This section– that covers the whole document – was reviewed and rated last in order that reviewers had the opportunity to gain an impression of the report as a complete work. In addition to assessing the coherence and style of the report, basic elements of an evaluation – such as the executive summary – were also considered under this section.

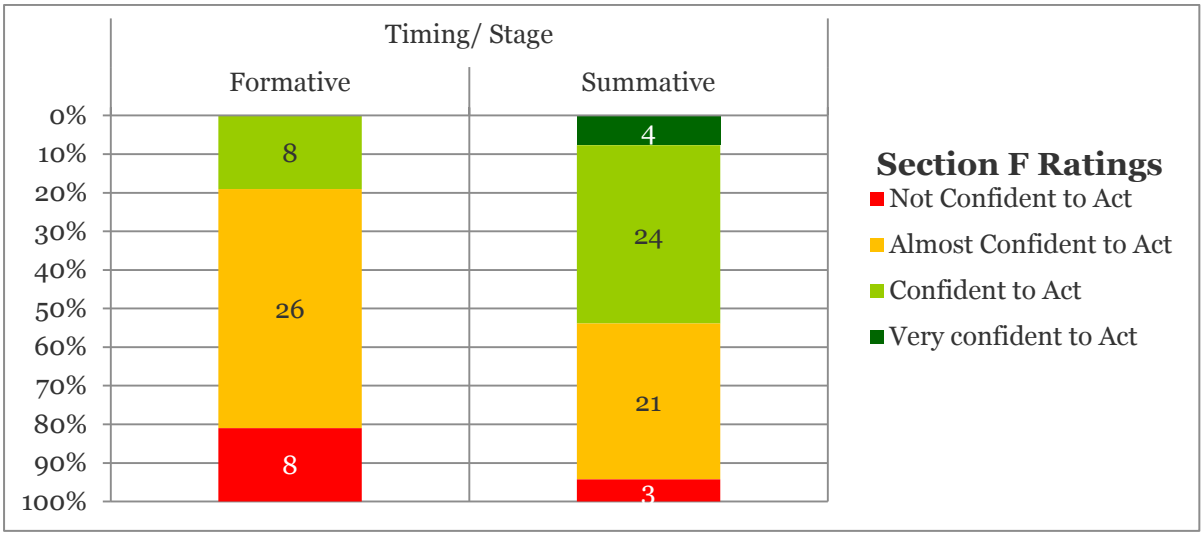
It quickly became clear that the majority of evaluation teams to not appear to have had sight of the UNICEF/UNEG minimum standards. Where possible, reviewers checked to ascertain whether these were referred to in the terms of reference: it was largely found that TORs did not draw attention to these standards.

To some extent, this part of the meta-evaluation was hampered by the fact that 66% of the reviewed reports did not have a copy of the TORs attached or embedded within them. This undoubtedly highlights a need for draft reports to be quality controlled for the basic elements before being accepted. In the 34% of reports that did have TOR attached these were often found to be unclear and imprecise.

The poor standard of many of the terms of reference is likely to be a significant contributing cause of many of the weaknesses observed within evaluation reports. For example, a good number of report shortcomings could be avoided with a pro-forma structure with and a detailed table of contents that includes the required annexes. TORs should also include guidance on the expected length of the report and the necessity of an executive summary.

Executive summaries are a required component of UNICEF evaluation reports. Although 86% of evaluations did include an executive summary, these were largely found to be weak and not fit for purpose. For instance, only 38% of reports had executive summaries that contained all of the basic elements, 48% of reports had executive summaries that were rated as unsatisfactory, and 14% of reports were missing executive summaries all together. Only 30% of reports included executive summaries that could confidently be used for decision making purposes: meaning that 70% of reports could not.

Figure 9: breakdown of report structure, logic and clarity ratings by timing/stage



There were, however, four examples of executive summaries that were universally *outstanding*. These were Colombia's Evaluation of Educacion en el riesgo de Minas, Guinea-Conakry/Guinea-Bissau's Evaluation of WASH Activities, Uganda's Evaluation of Programmes for Children Affected by Conflict, and Uzbekistan's Evaluation of Family and Child Support Services Project.

In addition to having strong executive summaries, reports that were rated *outstanding* for Section F differentiated themselves by including clear and accessible summaries of points throughout the report, by being logically structured, and by not using an excessive number of pages. Reports that were rated as *confident* tended to be those that were well structured. These appeared to be predominantly summative evaluations and those that considered outcome and impact level results.

Gender, human rights and equity

Ratings and feedback relating to the Human Rights Based Approach to Programming were integration into three of the review sections. The extent to which reports dealt appropriately with gender and equity issues were an integral part of these HRBAP questions. Whilst this approach successfully integrated the mainstreaming of HRBAP issues into how all relevant aspects of evaluation reports were rated, it proved to create a challenge in terms of drawing out lessons about gender, equity and human rights in an explicit manner.

Nevertheless, a number of strong patterns were found in the data. Just under half of evaluation reports (40) integrated *gender* considerations to some degree. This ranged from brief and scattered discussion of gender issues, through data disaggregation, to systematic and integrated applications of gender-based frameworks. Conversely, only seven reports dealt substantively with issues of *equity*[10]. Some reviewers noted that equity issues were included in the initial discussions of reports, but that these were not carried through into the methodology of analysis. For those evaluations that did attempt to deal with equity issues, the evaluators appeared to frequently rely on secondary data sources for their analysis.

Overall, only 30% reports were found to have methodologies that were appropriate for analysing gender and *human rights* issues identified in their scope. It was also noted that there appears to be a trend by which evaluations that were generally strong in other terms were rated poorly against the HRBAP questions, whereas some rights-focused evaluations rated very poorly overall. This trend can be seen in the disparity between the ratings of evaluation reports from different MTSP areas.

[10] Albania's Evaluation of the Impact of Social Assistance Mechanism on Reducing Child Poverty; Cambodia's Evaluation of Community-led Total Sanitation; Indonesia's EFA Mid-Decade Assessment; Nepal's Joint Evaluation of Education for All Sector Programme; Sudan's Evaluation of the Country Health and Nutrition Programme; Tanzania's Evaluation of COBET; Timor Leste's Evaluation of UNICEF Education Programme.

One set of evaluations that did stand out as being both strong on rights and strong overall were the various Child Friendly School evaluations. This might suggest that the existence of a strong rights-based guiding framework in the TOR can help to overcome some of the issues noted above. It was also found that the majority of evaluations in strongly rights-orientated areas – such as HIV/AIDS – were output level evaluations. As noted previously, these output level evaluations were rated significantly more poorly than higher level evaluations: possibly because they are conducted by technical specialists and researchers rather than people experienced in evaluation.

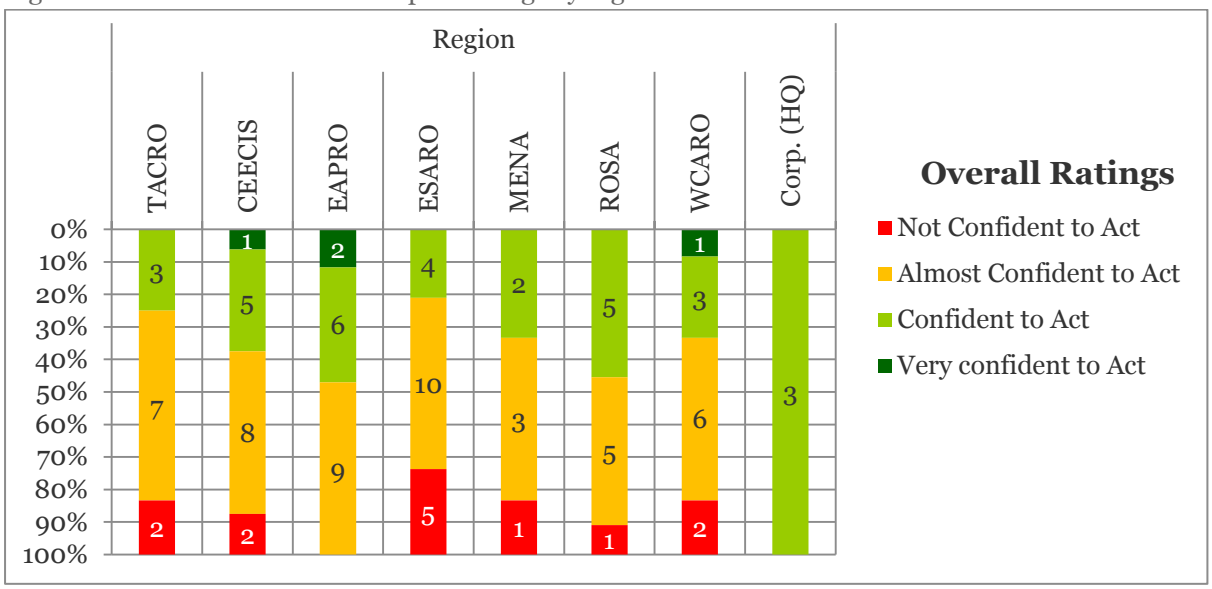
Regional findings

Whilst there were inevitable disparities across the quality of reports submitted by different regions, nearly all regions had at least one evaluation report that was rated *outstanding* in one or more of the assessed sections. The highest levels of satisfactory reports were concentrated in the Asian and European-based regional offices, with Timor Leste's Evaluation of the UNICEF Education Programme in the East Asia and the Pacific RO being the only report to receive a rating of *outstanding* across all sections and overall.

Although more reports were weaker overall in The Americas and Africa-based regional offices, Guinea-Conakry/Guinea-Bissau's Evaluation of WASH Activities in the West and Central Africa RO was rated as *outstanding* in all but one of the review sections, and *outstanding* overall. And, Colombia's Evaluation of Educacion en el riesgo de Minas in The Americas and Caribbean RO was rated *outstanding* in three sections, including for very under-addressed issues such as ethics.

It must also be recognised that the three global-level evaluations conducted by HQ Divisions were consistently rated as *confident* across all sections and overall. The Global Evaluation of Child Friendly Schools Programming, the Evaluation of DFID-UNICEF Programme of Cooperation, and the Global Evaluation of DevInfo thus serve as good example studies of reports that were able to maintain consistent performance across the different evaluation elements. There were, of course, examples of this consistency among selected reports from different regions. However, the combination of a small number of corporate evaluation reports and the consistent performance achieved across these would seem to illustrate the potential benefits of UNICEF focusing evaluation resources on delivering a smaller number of quality strategic evaluations.

Figure 10: breakdown of overall report ratings by region



Language differences

This meta-evaluation reviewed 76 reports written in English, 10 reports written in Spanish and 10 reports written in French. Some concern had been expressed by reviewers that non-English reports

were systemically weaker than equivalent English-language reports. Due to the relatively smaller number of non-English reports reviewed, our analysis considers the relative proportion of reports that are either satisfactory or unsatisfactory, rather than the four levels of ratings used elsewhere in this report.

Table 1: report ratings according to different languages

Report rating	All reports	French and Spanish reports	*French-language reports*	*Spanish-language reports*
Satisfactory	**36%**	**25%**	*30%*	*20%*
Unsatisfactory	**64%**	**75%**	*70%*	*80%*

(Columns 3 and 4 provide a breakdown of the data in column 2)

Based on this breakdown, it is apparent that French-language reports are largely consistent with the overall ratings of all reports, and that Spanish-language reports are somewhat weaker overall. This analysis does not show, but it is evident within the numbers, that there is a larger proportion of French and Spanish reports (60% and 70% respectively) rated as 'almost confident' compared to overall ratings.

Two explanations have been offered to explain any differences between the ratings of reports in different languages. Firstly, that reports in French and Spanish are primarily associated with the WCARO and TACRO regions. Secondly, that good practice guidance for evaluators is firstly provided in English and is not necessarily as comprehensive or 'original' (i.e. it is most often a translated version) in French and Spanish.

The first explanation – that differences can be accounted for by regional correlations – cannot be discounted. All ten Spanish-language reports were submitted to this review by TACRO region. There were only two other reports from this region, and thus the ratings for TACRO region correlate strongly with those for Spanish-language reports. Similarly, based on data about Spanish-language reports, the second explanation – that guidance is weaker – also cannot be discounted because the group of reports from TACRO rated overall as the second-weakest regional-block.

French-language reports were submitted by three regions: ESARO, MENA and WCARO. Of these, 70% of French-language reports (seven reports) were from WCARO region. WCARO also submitted four English-language reports. Ratings for WCARO reports are the same as global ratings, 36% being satisfactory and 64% unsatisfactory. Within the region, French-language reports (43% satisfactory with no outstanding examples) were rated more highly than English-language reports (25% satisfactory with one outstanding example).

These findings suggest that the data is currently insufficient to draw any firm conclusions regarding language differences and that further investigation of the question may be required. Such investigation may also wish to consider any differences between English and Russian reports in CEE/CIS and English and Arabic reports in MENA.

Examples of interesting and innovative practices

During the review process, a number of notable practices were observed in each of the UNICEF regions. These were noted in the final section of the review tool and reported back to evaluation managers. The highlights were:

Table 2: notable practices encountered in evaluation

TACRO	• Prevalent use of participatory methodologies • Clear design and explanation of evaluation frameworks in accessible language • Explicit attempts to avoid bias in documentary analysis • Including statements on ethical standards • Use of case studies to illustrate analytical points • Strong use of qualitative analysis to process data

CEE/CIS RO	• Use of internationally agreed indicators and links to global policy context • Clear concise writing styles and accessible language that enhance credibility • Integrated use of cost analysis • Including UNDAF results frameworks, baselines and updated indicators in annexes • Systematic causal reasoning and exploration of unexpected findings • Use of the Chabot Matrix to cross tabulate child-friendly and quality issues in analysing CFS
EAPRO	• Discussion on ethics and bias in the use of translation • Use of counterfactuals in simple and manageable approaches • Methodologies developed through consultation with stakeholders • Create large databases of evidence from multiple sources • Use of detailed project timelines
ESARO	• Ethics included as integral part of a methodology • Critique of the definition of impact that was being used • Observational exercised on the use of ICT investments • Stories in pictures and carefully selected pictures to illustrate report findings • Participatory evaluation design • Strong critique of project M&E systems • Participatory observation and field visits
MENA RO	• Use of detailed research questions linked to the evaluation objectives • Inclusion of the results chain diagram • Use of scenarios to present recommendations
ROSA	• Analysis of programme compliance with Paris Principles on aid effectiveness • Evaluation approach and methodology was included as an integral part of the project design • Child centred methods to understand the views of children • Comparable evaluations run in multiple countries
WCARO	• Strong use of case studies • Prevalent use of mixed methods approaches • Comparison of assessed project with comparable benchmark projects • Systematic review of the literature
Global/ corporate	• Flagging good practice examples across a range of different contexts • Use of hierarchical linear modelling • Strong focus on data presentation and usability

Table 3: notable thematic practices encountered

TACRO	• Creating community and family dialogue around the value of education • Use of an action without harm framework to create security and protection of child rights activities • Focusing on strategic education partnerships
CEE/CIS RO	• Supporting specialised sections of justice sector institutions for children and youth • Integrated approaches to maternal and child health in place of vertical approaches
EAPRO	• Supporting Juvenile Justice Committees to be effective • Adapting existing national systems to deliver new types of services for children • Creating conducive institutional environment for basic education training to have impact
ESARO	• Low-cost low-tech local latrine solutions to reduce flies and mosquitoes • Working with major philanthropy (the Hunter Foundation) • Using ECD as an entry point for child rights • Focusing on the self-esteem and confidence of peer educators
MENA RO	• Making long term training commitments across sectors in order to reinforce gains

ROSA	• Using evaluation as an integral part of knowledge management to adapt a project to 'what works' • Exploring different gender needs from IEC materials for HIV/AIDS • Flash reports in the education sector that give a biannual snapshot of education data • Welcome to School project influencing national policy
WCARO	• Door-to-door visits as a mechanism for warning about and reduce cholera deaths • Public Declarations by communities
Global/ corporate	• The Child Friendly Schools framework as sufficiently clear and sufficiently flexible for a wide range of contexts

Conclusions

The following conclusions were developed by analysing the findings for trends in underlying factors that contributed to the performance of evaluation reports. This analysis was grounded in the concept of 'confidence', as articulated in the UNICEF Global Evaluation Report Oversight System. Conclusions were developed pertaining to both the quality of reviewed evaluation reports and the review methodology itself.

The quality of evaluation reports

Evaluation reports benefit from having access to relevant and well-developed international frameworks

The *MTSP* and *purpose* analyses clearly reveal a tendency for evaluations of education and humanitarian objects to have reports of better quality than their contemporaries[11]. Our investigation into this trend suggests that these two areas benefit from having well-known, mature and contextually-adaptable frameworks. In the case of education these are the Child Friendly Schools framework and the Education for All framework; in the case of humanitarian evaluations it is the Core Commitments to Children.

Being able to use these frameworks to guide an evaluation report appears to help evaluators to both better organise the report as a document, and more consistently structure the analysis across findings, conclusions and recommendations. Furthermore, by nature of being rights-based these frameworks appear to help attenuate some of the disparity between good quality handling of rights and evaluation issues that are discussed in the next conclusion.

A disjuncture exists between successful evaluation and strong integration of rights

There would seem to be a complex and multi-faceted dynamic around an apparent disparity between reports that respond well to rights issues and reports that rate well overall. Our investigation into the causes behind the poor ratings of output evaluations compared to outcome and impact evaluations, and the poor rating of formative evaluation compared to summative evaluations, found that 70% of the output evaluations were also formative evaluations.

Whilst this observation is not entirely unexpected (formative evaluations are naturally more likely to not yet have access to data on higher-level results), it combines with the observation made previously of a strong correlation between rights-orientated evaluations and output evaluations. At this point, we conjecture that there may be a link to the observations by reviewers that: a) project and programme results frameworks cited in reports tended to be generally weak and have unclear indicators; and b) rights evaluations tended to make use of technical specialists rather than evaluators.

Based on this evidence, we propose that there may be a dynamic by which rights-based projects and programmes do not have strong line-of-sight to clear, relevant and accepted measures of expected outcomes and impacts. As a consequence, there is a greater tendency to commission lower-level outputs evaluations that appear to be smaller and less complex than outcome or impact evaluations. Less complex evaluations attract lower levels of evaluation-resource and attention, and thus evaluation managers rely instead on the technical specialists familiar with their project or programme because of previous situation analyses or research work.

This situation may be compounded by a lack of easily accessible evaluators with rights expertise. The result of this dynamic is inevitably a poorer quality of rights-orientated evaluation reports. Such a hypothesis is consistent with the observation that complex evaluations that *do* invest heavily in

[11] This might also be caused by reviewer bias; although this explanation has been discounted as a significant proportion of reviewers have a rights background.

evaluation skills (multi-sector, cross-cutting and corporate) all rate as consistently strong across all sections of the review. From the evidence available to this review, it would appear that there are two central drivers to this whole dynamic:

1. Fragmentation of skills. *It would appear to be the case that Country Offices often have to select between either consultants with rights knowledge or consultants with evaluation skills. Where one of these specialities is selected, the other aspect inevitably suffers in terms of the quality of overall reporting. It is not clear why this is the case in terms of either the individuals recruited or the evaluation teams that are put together, but it could be speculated that some form of resource-constraint might be a contributing factor.*
2. Unmet needs for strong mainstreaming frameworks. *As discussed previously, the framework for mainstreaming HRBAP, gender and equity in the evaluation function appears to be both underdeveloped and under-communicated. Evaluations that do perform well in terms of both rights and overall are those that have access to a strong rights-based framework such as CFS (see the conclusion above).*

The evaluation function is not delivering consistent contributions to upstream knowledge management

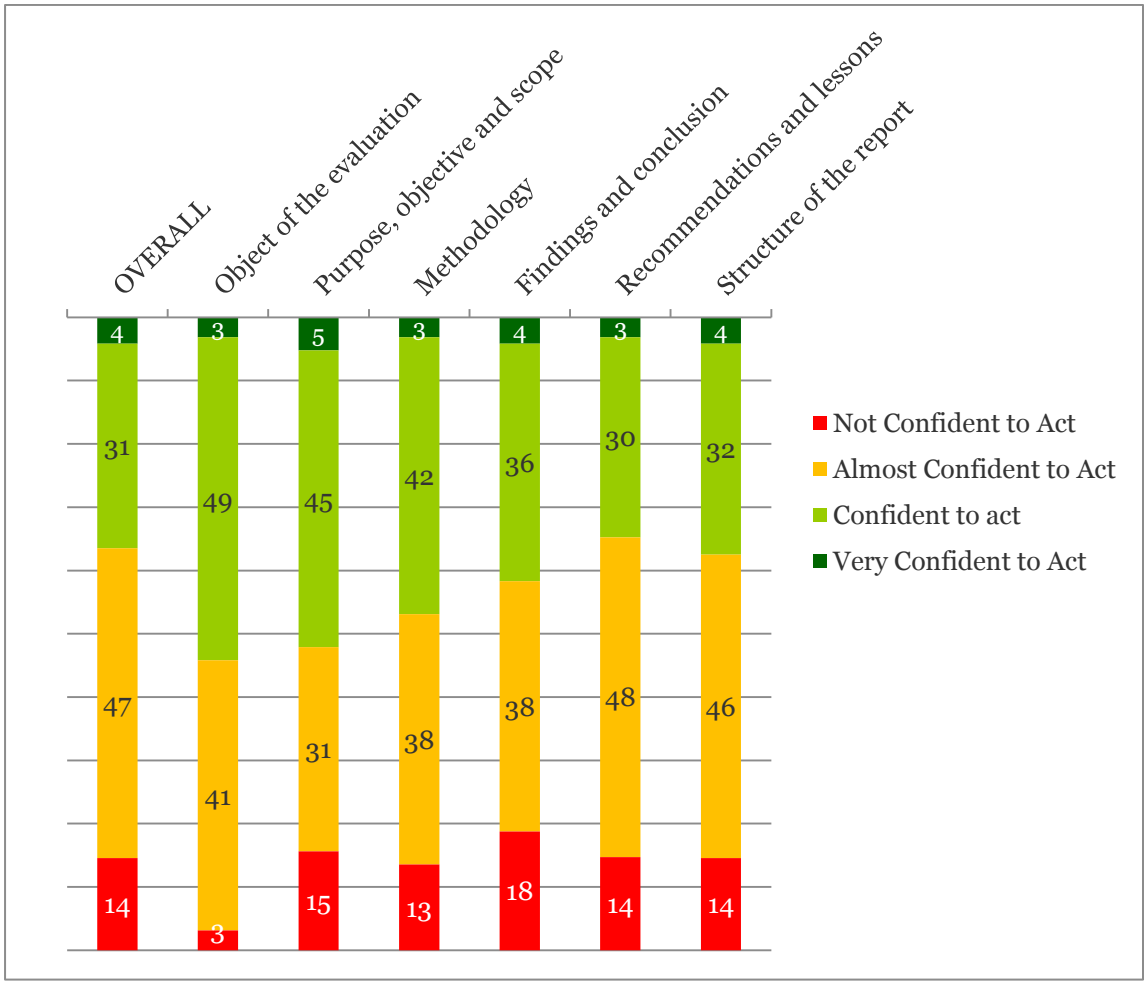
The *purpose* and *MTSP* analyses both found that policy evaluation reports and organisational performance evaluation reports are weak areas. For instance, none of the policy evaluation reports reviewed was rated as satisfactory in relation to recommendations: essential to supporting better positioning of UNICEF in an upstream context. Four out of five organisational performance evaluation reports were rated as unsatisfactory overall even though these can make an essential contribution to making the organisation fit-for-purpose in an upstream context. Thus, from the perspective of UNICEF's upstream ambitions, the current performance of the evaluation function is likely to be of some concern.

Robust and transparent analysis of data is a problem

All the different ways of breaking down the rating data reveal one consistent trend: evaluation reports are stronger in the initial sections of the review, with performance gradually deteriorating over the span of the report. To some extent this trend may be explained by natural attrition: latter sections relying on the performance of preceding sections are in some sense starting from a 'loaded' baseline. However, the example of the three corporate evaluations demonstrate that not only is it possible for reports to deliver consistently satisfactory ratings across all aspects of the review, but that it is also possible for a report that struggles a little on methodology to deliver strong findings, conclusions and recommendations.

Thus the central issue appears to be that evaluators are far clearer about the theory of evaluation (purpose, objectives, methodology, data collection) than the processing and analysis of data that is generated. Reviewers noted a frequent breakdown in the logical threads that should link data to findings to conclusions to recommendations. This appeared to be caused primarily by inadequate, opaque or inappropriate analysis of the information presented. Reports included raw data in the findings section, tended to jump to conclusions without explaining how, and gave recommendations based on the personal interpretation of the evaluator. Each time analysis was inadequate between different elements of the evaluation report, so subsequent sections would rate more poorly than their immediate predecessors: feasibly explaining the meta-trend of decreasing performance.

Figure 11: breakdown of ratings by section



Weak terms of reference are contributing to poor report quality

As discussed extensively in the previous section, problems in the terms of reference for evaluations were found to be a heavy contributor of later quality issues in the report. Reports tended to 'build' off of the TOR as a starting point in terms of the evaluation purpose and framework, so better TORs inevitably resulted in better reports. This reemphasises the value of conducting basic checks and quality assurance on TORs, ensuring that each TOR gives evaluators sight of the minimum standards, and using the TOR to articulate very clearly the purpose of the evaluation.

Fundamental misunderstandings of recommendations and lessons learned prevail

Most reports attempt to include recommendations and lessons learned sections[12]. Few of these are successful. Recommendations were often found to be disconnected from the preceding sections, drawing on the personal knowledge or opinions of the evaluator(s). It also appears to be enormously difficult for reports to consistently strike a good balance between actionable specificity and avoiding long lists of over-detailed recommendations.

Lessons learned generally perform even more poorly overall. In the vast majority of cases it became immediately apparent that the purpose of lessons learned was not well understood. Many reports included project or programme-specific operational issues under lessons learned, others referred to the constraints of the evaluation itself. Indeed, the prevalence of misunderstanding of the lessons learned element of reports might suggest that it is a central candidate for explicit efforts to raise the awareness of both evaluation managers and evaluation teams about what lessons learned are.

[12] Both are integral parts of an evaluation report according to UNEG Standard 3.16.

The meta-evaluation process, methodology and review tool

In addition to the evaluation report quality conclusions stated above, the meta-evaluation process generated two central conclusions around the process, methodology and review tool of the meta-evaluation itself (these are drawn from the analysis presented in Annex 6).

The qualitative approach is a viable and useful way forward for the UNICEF Global Evaluation Report Oversight System

As discussed at some length within the findings section, the experience of the meta-evaluation has found the qualitative approach to not only have been capable of generating analysis equivalent to or richer than the previous quantified methodology, but that it also enabled reviewers to provide more useful, accurate and constructive feedback to UNICEF managers in a range of contexts. It will be interesting to test this conclusion – based largely on reviewer feedback – against the feedback from review-recipients.

There is work to be done in supporting evaluation managers to mainstream human rights, gender and equity

The challenge of locating intelligent assessment of human rights, gender and equity issues within the evaluation review format, and the generally weak performance of these areas overall, suggests that there is a real need for stronger guidance and support for evaluation managers on these issues. It maybe that this takes the form of a coherent organising framework for HRBAP within the evaluation function, but the central conclusion remains that there are some profound challenges in the interaction between HRBAP and robust evaluation that are unlikely to be addressed through simply revising weaknesses in the review tool.

Lessons Learned

These three lessons learned have been generated through the analysis of the core-evaluation team. They have been adapted based on subsequent dialogue with and contributions from the UNICEF Evaluation Office and comments from regional offices.

There is great value to be gained from blending the skills of evaluation teams

UNICEF is institutionally committed to delivering on human rights and delivering results and accountability through the use of robust evaluations. From the evidence generated through this meta-evaluation it would appear that evaluations which rely on either technical specialists or evaluation experts tend not to be able to satisfactorily deliver on both of these commitments. However, complex rights-orientated evaluations have been delivered successfully where both evaluation skills and sector-knowledge has been present on the team. With an apparent shortfall in both evaluators with rights-knowledge and rights-specialists with evaluation skills, it would seem that value could be delivered from creating evaluation teams that mix these two skill-sets as an alternative to the more traditional international/local knowledge blend.

Developing strong international frameworks provides a platform for stronger, more rights-orientated and more useful evaluation

Sectors that have historically invested in developing widely-accepted and understood rights based frameworks benefit from having consistently more useful evaluations as a result. To some extent this is because they provide an organised structure and logic for analysing data. However, it would also appear that frameworks such as Child Friendly Schools and Core Commitments to Children empower evaluators to better manage rights issues within their evaluations. Stronger, more useful evaluations contribute to enhancing knowledge across these sectors and thus strengthening the frameworks themselves: thereby contributing to the creation of a 'virtuous spiral'.

Co-designing the methodology and investing in the development stage of evaluation-quality-reviews delivers a strong return in performance

The conclusion that the qualitative approach is a viable and useful direction for UNICEF is based on more than the methodology or tool itself. The elements of the meta-evaluation that have proven to be

most valuable – the organisation of the sections, the definition of a confidence test, the inclusion of constructive feedback, the dynamic aggregation of qualitative ratings – were all made possible by the front-loading of investment in this exercise.

Misunderstandings that could have become a problem at the analysis stage were eradicated early on through face-to-face working between UNICEF and IOD PARC teams. This also had the benefit of attenuating different interpretations of ratings by reviewers (further reduced through the quality assurance processes described in the methodology). Reviewers themselves had a chance to test and to help refine the review tool before it was finalised and deployed. This had the benefit of working through many possible scenarios and ultimately contributing to a more universally usable tool that can be reused with little substantive adaptation.

Recommendations

The following recommendations have been generated through the analysis of the core-evaluation team. They have been adapted based on subsequent dialogue with and contributions from the UNICEF Evaluation Office and comments from regional offices.

Recommendations for UNICEF Evaluation Office

Focus on delivering a strategy for more consistently high-quality terms of reference within the decentralised evaluation function

The significance of terms of reference in the current performance of evaluation reports has been a consistent presence in all stages of this meta-evaluation. This strongly suggests that it is a priority area for action that could deliver wide-ranging gains. Options that could be explored include the communication of a basic pro-forma terms of reference for evaluation that includes a generic table of contents (this information already exists in UNICEF standards, but these do not appear to have been given to most evaluation teams).

Other options might include using the comparative advantages of the Evaluation Office to support quality assurance of terms of reference by regional offices, be that through helpdesk facilities or other means. Whichever solutions are developed will inevitability need to be flexible enough to respond to the wide range of contexts encountered by evaluations in UNICEF. Although the issue has not been discussed at length in this report (because the data suggests that it is not yet a priority), mechanisms to enhance the quality of terms of reference will need to be flexible enough to cope with increasing numbers of joint and country-led evaluations.

Continue development of the qualitative approach to deliver the Global Evaluation Report Oversight System

Based on the experience of this meta-evaluation, of the previous global evaluation of report quality, and of regional meta-evaluations of report quality it is recommended that the qualitative assessment approach offers substantive benefits over the previous quantified approach and should be continued. Furthermore, the findings of this report would appear to suggest that the flexibility of the review tool makes it more easily adaptable in a relevant way to different contexts and levels of evaluation. Further work to formalise the approaches developed here is thus likely to be of benefit more widely within the organisation than quality assuring the Global Evaluation Database.

Develop and communicate an integrated HRBAP mainstreaming strategy for the evaluation function

The review tool used for this evaluation needs to be enhanced in terms of human rights, gender and equity assessment. However, this needs to be part of a more systemic strengthening of HRBAP within the evaluation function. A strategy and the tools for doing so are likely to have the most profound

benefit if they are communicated effectively to evaluation managers and integrated as a central part of terms of reference for future evaluations.

Invest in clarifying results frameworks for challenging thematic areas, work to evaluate these at higher levels, and contribute this knowledge to international attempts to develop coherent frameworks

The value to evaluation report quality of having coherent international frameworks has been highlighted several times in this report, as has the cost of having weak or non-existent results frameworks. Not all cutting-edge themes and sectors benefit from a legacy of clear logic models or outcome and impact level indicators. Where this is the case, there appears to be a tendency to focus instead on output level evaluations that consistently deliver little in the way of real value.

Thus, it might be a better approach to focus instead on doing a lesser number of higher-level evaluations in challenging areas and using these as a platform to invest in clarifying the results framework. Where resources are constrained, or where innovative models are developed within a country, there is a clear role for regional offices to help address these issues through developing multi-country collaboration.

Revisit how the requirements for recommendations and lessons learned sections are communicated to evaluators and managers, particularly in relation to policy-level evaluation

The systemic weakness of recommendations and lessons learned across the assessed evaluations would seem to represent a significant barrier to the evaluation function delivering optimum value. This is particularly pronounced in relation to policy-level evaluations: an area where the value of lessons learned would be of greatest value to UNICEF's increasing focus on upstream working. The Evaluation Office has developed an increasing array of communications tools to reach out to the decentralised evaluation function (such as MyM&E). It would appear to be timely to consider how these can target the capacity of policy evaluation managers and evaluators to deliver better recommendations and correctly-identified lessons learned.

Recommendations for Regional Offices and Country Offices

Recommendations for regional and country offices have been combined because they largely require joint working in order to deliver successful evaluation reports.

Cooperate on delivering basic quality assurance at the TOR stage and draft report stage of evaluations

Monitoring and evaluation resources are often highly stretched within country offices. However, systematically applying even the most basic quality assurance checks to make sure that terms of reference and evaluation reports contain the essential elements can already make a substantial difference to the quality of reports that have been reviewed here. Where country level resources are stretched or in transition, regional and country offices should explore options for delivering these basic checks through systematically reviewing evaluation TORS and draft reports at the regional level.

Attempt to focus on fewer and better evaluations that deliver strategic priorities

Reducing the number of evaluations across complex country programmes with multiple donors is not always easy. However, the evidence presented here suggests that where it can be done then there may be significant gains in terms of evaluation report quality. The aid effectiveness principles of harmonisation, alignment, managing for results and national ownership can all be invoked to get donor support for using the IMEP to prioritise and focus on strategically important evaluations. Initially, these may require methodologies that can differentiate the contributions of different donors in order to develop buy-in, but transitioning to strategic evaluation will also open up opportunities to engage in more joint and country-led evaluations.

Create cross-UNICEF pollination around upstream evaluations, and explore the options for multi-country approaches

The multi-country Child Friendly Schools evaluations reviewed in this meta-evaluation highlight the opportunities of the critical mass strategy to building internationally-influential knowledge that both enhances and learns from individual country experiences. Within regions, delivering at least one example of a successful and relevant multi-country evaluation is likely to enhance the wider evaluation function in these countries at the same time as influencing policy and demonstrating the value-added of the organisation. If this can be linked to the international agenda then it also provides a strong platform for the cross-pollination of knowledge within UNICEF.

Deliver extra support to country-led evaluations

The evidence developed by this meta-evaluation is insufficient to identify the reasons why country-led evaluations rate more weakly than those managed by UNICEF of UN partners. However, the aid-effectiveness principles of the Paris Declaration and Accra Agenda for Action, to which UNICEF is committed, are increasing the prevalence of such evaluations. Concomitantly, the UNICEF Programme Performance Assessments of middle income countries (synthesis forthcoming) have found that results based management is a major entry point for strategic engagement between UNICEF and national governments. Thus, it is recommended that Country and Regional Offices consider investigating the challenges experienced by country-led evaluations within their contexts and responding accordingly.

Potential areas of focus for improving evaluation report quality in specific regions

Based on our reading of the evidence, the following suggestions are offered as to possible focus areas for individual regions to move forward on enhancing the quality of evaluation reports.

Table 4: regional recommendations

TACRO	Focus on enhancing the processes for developing and communicating recommendations that are used by evaluators • This is based on your Section E ratings
CEE/CIS	Focus on maximising experience with rights issues by increasing the involvement of evaluators and tackling higher-level results • This is based on qualitative feedback analysis of rights evaluations
EAPRO	Focus on more consistently structuring evaluation reports and eradicate unsatisfactory conclusions sections • This is based on your Section D and F ratings
ESARO	Focus on enhancing the essential ingredients of a good evaluation through improving the quality of terms of reference • This is based on your overall performance ratings
MENA	Invest in developing access to a wider cohort of evaluators with relevant contextual knowledge and experience • This is based on your Section A ratings
ROSA	Focus on quality assuring the analytical capability of evaluators by requiring detailed methodologies to be provided in inception reports (including analytical frameworks and processes) • This is based on your Section C and D ratings
WCARO	Explore options for sharing and learning from regional examples that are outstanding, particularly in relation to mixed methods • This is based on your overall ratings and best practice feedback
Corporate	Focus on sharing lessons for consistent evaluation with regions and try to maintain this consistency with a higher rate of knowledge generation • This is based on your consistent section ratings

Annexes

Annex 1: Terms of Reference

UNICEF Global Evaluation Quality Oversight System

Background

UNICEF Evaluation Office (EO) put in place an Evaluation Quality Oversight System to monitor the impact of efforts to strengthen the UNICEF evaluation function globally. The system consists of rating evaluation reports commissioned by UNICEF Country Offices (CO), Regional Offices and HQ divisions against the UNEG/UNICEF Evaluation Report Standards. Reports meeting satisfactory rating are made available in the UNICEF Global Evaluation Database (GED).

The database was created on the UNICEF Intranet in 2001. In June 2002, it was made available to the public on the UNICEF Internet website. It currently contains around 3300 records.

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

Expected results

The selected institutions will review all evaluation reports in English, French and Spanish received by the EO, rate them against UNEG/UNICEF standards, write an executive feedback to be sent to the CO concerned, and make an analysis of trends, key weaknesses and strengths of UNICEF-supported evaluation report.

Expected deliverables

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

A. Executive feedback template

Develop draft executive feedback templates and finalize them based on Evaluation Office comments. The Executive feedback should be one to two pages, and should focus on strengths, weaknesses and recommendations to improve quality of future evaluations reports**.**

B. **Evaluation reports reviewed, rated and executive feedback sent**

Review and rate Evaluation reports received in English, French and Spanish against UNEG/UNICEF standards (annex 1) using the UNICEF/UNEG Evaluation Report Review Matrix (annex 2), and send ad-hoc executive feedback as well as the Evaluation Report Review Matrix filled (with a qualitative rating such as “satisfactory”, “good”, etc.) to the concerned UNICEF Office, through the UNICEF Evaluation Office, highlighting strengths, weaknesses and recommendations to improve the quality of future evaluations reports**.** The completed Evaluation Report Review Matrix with the actual numerical rating should go the Evaluation Office only.

The estimated total number of evaluation reports to be reviewed will be approximately 300, of which about 80% are in English, 15% in French and 5% in Spanish.

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

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

After reviewing the reports on a yearly basis, produce a Global analysis (no longer than 20 pages) of trends, key weaknesses and strengths of reports reviewed, including lessons learned and good practices on Evaluation reports, and actionable conclusions and recommendations to improve the Evaluation Quality Oversight System as well as the quality of Evaluation reports. The outline of the global analysis will be developed in consultation with UNICEF Evaluation Office. This will enhance COs capacity to ensure future reports meet the criteria and standards as set out by UNCEF/UNEG. As evaluation reports are actively shared internally and externally, it is important that reports shared are of high standards.

Management of the system

The Evaluation Quality Oversight System will be managed by the Senior Evaluation Specialist, Systemic Strengthening, with the support of the Knowledge Management (KM) Specialist.

The selected institution(s) will appoint a project manager who will ensure consistency of rating, quality and timely delivery of expected products, and overall coordination with UNICEF Evaluation Office. The project manager will also provide an update on a monthly basis, which will include a tracking matrix highlighting the status of reviews, ratings and executive feedback.

Qualifications

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

Annex 2: The Review Team

The Core Evaluation Quality Team was made up of ten specialist consultants in Monitoring and Evaluation. They were led by **Sadie Watson**, who was responsible for high level quality assurance, **Hatty Dinsmore**, responsible for overall project management, and **Joseph Barnes**, lead author.

Sadie Watson is a Director with over 10 years' experience as a monitoring and evaluation consultant with a specific focus on social development processes and methods. She has a strong background in monitoring, evaluation and organisational reviews and how these processes contribute to organisational learning and improving the way organisations work.

Hatty Dinsmore is a consultant with core skills and experience in project and financial management in international development. Her experinece of living and working in East Africa brings further depth and insight into what constitutes effective and efficient project management and brings a significant appreciation for how to build and manage networks of resources.

Joseph Barnes is a senior consultant with a strong grounding in results based management, international development policy, ethics, and disaster reduction. Joseph combines field experience and applied practice in a variety of programming challenges to his consulting work which focuses primarily on organisational analysis and translating strategic programme design into results-based operations.

Annalize Struwig is a principal consultant and has over 20 years' experience as a development consultant and donor advisor specialising in M&E, health systems research and HIV/AIDS. As a consultant, Annalize has worked for a cross section of development agencies, including DFID, the EC, WTO and Irish Aid. Her work for these agencies includes specialist inputs on M&E, as well as the evaluation of projects, sectoral and inter-sectoral programmes and country programmes.

Frank Noij is an associate with 23 years of experience in evaluation and results based management. He has led and participated in evaluations on country programme, thematic, program and project level and has a track record on quality assurance in evaluation. He has worked on organisational capacity development for results based management and supported organisations in adapting their M&E systems to the requirements of a human rights-based approach to development programming.

Julia Betts is a principal consultant specialising in M&E, education and social development. She has over 15 years' experience applying learning and theory to development in practice, working at levels from the grass roots to the international donor agency (DFID). Building on strong analytical and strategic thinking skills Julia's main consulting focus is on evaluation and performance assessment, social analysis and social impact assessment, equity, rights and inclusion (in particular gender analysis, planning and mainstreaming).

Pierre Robert is an associate with over 20 years' experience as an international policy research expert with detailed knowledge of civil society organisations worldwide and strong experience in M&E. Pierre has practical experience of social impact assessments, institutional reform, gender, democratisation, access to justice and social development processes, conflict prevention, and expertise on international organisations, project management and human rights, with a strong background in political and social issues in Asia and Africa.

Ronnie MacPherson is a senior consultant with over 10 years of experience in international development. His work focuses on bringing together organisational planning & development, monitoring & evaluation, information technology, and knowledge management. Ronnie has worked extensively with donors, policymakers, international development NGOs and UK-based domestic charities, supporting and delivering organisational and strategic development and facilitating collaborative work between diverse organisations.

Francis Watkins is a core associate and social development specialist with over 18 years' experience working bilateral and multilateral agencies, governments and international NGOs. Francis has extensive experience in a variety of sectors and has a wide range of experience leading and working in teams in project and programme design, development and implementation of sectoral work, and monitoring and evaluation systems.

Sheelagh O'Reilly is a Director with over 20 years' experience working with international organisations engaged in development and has a wealth of experience in M&E, evaluation management and development effectiveness. Sheelagh has a broad background within the natural resources arena covering industrial management and research, academia and field project implementation linked to policy development. This experience is linked to an understanding of international legal regimes including those associated with environment, human rights, trade (including IPR) and corporate social responsibility.

Each consultant brought a depth of evaluation experience in international development and the team was balanced in a wide range of specialisms including; gender, social exclusion, natural resources, child rights, humanitarian development, rural and social development, health, HIV/AIDS, education, environment and human rights. The team included native speakers in English, Spanish and French which ensured high standards of reviewing in each language.

Annex 3: List of assessed reports

Country	Sequence number	Title	Overall rating
Albania	2009/006	Evaluation of CPU Intervention in the Framework of Social Welfare System Reform and Decentralization of Social Services	Not Confident to Act
Albania	2009/011	Assessment of Juvenile Justice Reform Achievement in Albania	Almost Confident to Act
Armenia	2009/005	Evaluation of Inclusive Education Policies and Programmes in Armenia	Confident to act
Armenia	2009/007	An assessment of Armenia's Child Friendly School Pilot Projects and CFS standards for UNICEF Armenia	Almost Confident to Act
Azerbaijan	2009/001	Evaluation of the United Nations Development Assistance Framework 2005-2010	Confident to act
Bangladesh	2009/006	Evaluation of PRECISE: a comprehensive Child Injury Program in Bangladesh	Confident to act
Benin	2008/007	Evaluation et actualisation du Paquet Educatif Essentiel pour l'accélération de la scolarisation des filles au Bénin	Almost Confident to Act
Bosnia & Herzegovina	2008/015	External Evaluation of the "Child-Friendly Schools" Project 2002-2007	Confident to act
Burkina Faso	2009/006	Evaluation des investissements en infrastructures réalisée par l'UNICEF au Burkina Faso (2006-2008)	Confident to act
Cambodia	2009/001	Law Enforcement against Sexual Abuse, Sexual Exploitation & Trafficking of Children Project - Phase 3 Assessment	Almost Confident to Act
Cambodia	2009/002	An Evaluation of the Anti Trafficking and Reintegration Programme of the Ministry of Social Affairs, Veterans and Youth Rehabilitation, Cambodia	Almost Confident to Act
Cambodia	2009/003	Community-Led Total Sanitation	Confident to act
Cambodia	2009/005	Assessment of the HIV/Reproductive Health Programme - Health for Future Work	Almost Confident to Act
Cambodia	2009/006	Impact Assessment of Basic Education Program (EBEP) Supported Training	Almost Confident to Act
Cameroon	2010/001	Evaluation of the parental education for the development of the young child strategy in the Adamawa region	Not Confident to Act
Chile	2008/008	Evaluación del apoyo de UNICEF a la educación municipal en la comuna de Conchalí	Almost Confident to Act
Colombia	2009/002	Evaluation of the 'Return to Happiness' methodology as a strategy for psychosocial recovery and as a component of the strategy for preventing the recruitment of children and adolescents by illegal armed groups	Not Confident to Act
Colombia	2008/004	Proyecto "Educación en el riesgo de Minas (ERM) y Asistencia a Victimas en los departamentos de Cauca, Chocó, Nariño, y la región de la Mojana/Sur de Bolivar", Colombia	Confident to act
Colombia	2008/010	Proyecto de Saneamiento Básico Ambientas (SBA) y Tratamiento de Aguas Residuales en Zonas Rurales del Departamento de Caldas: Evaluación Ex-Post del Proyecto: Informe Ejecutivo (An Ex-Post Evaluation of the Project to Improve Family Sanitary Units in the Rural Area of Caldas)	Almost Confident to Act
Colombia	2009/003	Evaluación Del Programa "Escuela Busca al Niño"	Almost Confident to Act
Colombia	2008/003	La Garantía y La Protección de Derechos De La Infancia, La Adolescenia y La Juventud En Los Planes De Desarrollo de Los Departmentos y Los Municipios de Colombia, 2008-2011- Level of Inclusion of Adolescents and Youths in Territorial Development Plans	Confident to act
Colombia	2009/004	Evaluacion Manuales de Convivencia Escolar: Aplicacion de contendidos y Participacion Activa De Ninos, Ninas y Adolescented En Su Definicion - Magdalena Boyaca y Soacha - Colombia	Almost Confident to Act
Corporate (HQ)	2009/001	Children and the 2004 Indian Ocean Tsunami: Evaluation of UNICEF's Response in Indonesia, Sri Lanka and Maldives (2005-2008) – Overall Synthesis Report	Confident to act
Djibouti	2009/001	Rapport De L'évaluation Externe De La Phase Pilote Du Projet De Prise En Charge Des Orphelins Et Autres Enfants Vulnérables En République De Djibouti	Almost Confident to Act
Dominican	2010/001	Evaluación de Impacto de la Estrategia de Comunicación y	Almost Confident

Country	Sequence number	Title	Overall rating
Republic		Movilización Social "Sal Yodada.....o Nada"	to Act
El Salvador	2010/001	Evaluación Del Programa De Seguridad Ciudadana: Prevención del delito y de la violencia social desde la perspectiva de los beneficiarias	Almost Confident to Act
Ethiopia	2009/001	Evaluation of the EthioInfo Utilization in Ethiopia	Almost Confident to Act
Gambia	2009/001	Evaluation of the Parental Education Programme at LRR, CRR, URR, The Gambia	Not Confident to Act
Gambia	2009/002	Process Evaluation Of The Joint Government, Unicef & Tostan Pilot Project In The Gambia	Almost Confident to Act
Georgia	2009-800	UNICEF's Response to Georgia Crisis: Real Time Evaluation	Confident to act
Global	2009/009	Child Friendly Schools Programming: Global Evaluation Report	Confident to act
Global	2009/008	Evaluation of DFID-UNICEF Programme of Cooperation, Investing in Humanitarian Action, Phase III (2006 – 2009)	Confident to act
Global	2008/818	GLOBAL EVALUATION OF The Development Information System (DEVINFO)	Confident to act
Guinea	2009/002	l'Evaluation des résultats des 12 Associations de Services Financiers (ASF) pour le développement Communautaire dans la zone de Kissidougou	Almost Confident to Act
Guinea-Bissau	2010/001	Evaluation of Cholera Surveillance System in Guinea-Bissau	Almost Confident to Act
Guinea-Bissau	2010/002	Evaluation of WASH activities undertaken to prevent and control cholera outbreak in Guinea Conakry and Guinea-Bissau – A systematic Review	Very Confident to Act
India	2007/021	Assessment of Effectiveness of the Avian Influenza Communication Interventions in 4 Districts	Almost Confident to Act
India	2008/021	Evaluation of the Child Reporters Initiative (CRI)	Almost Confident to Act
India	2008/034	Assessment of Effectiveness of IEC Material at Integrated Counselling and Testing Centres	Almost Confident to Act
India	2009/041	Effectiveness of IEC materials at Red Ribbon Clubs for HIV Prevention	Almost Confident to Act
Indonesia	2007/010	EFA Mid Decade Assessment Indonesia	Confident to act
Indonesia	2008/010	Assessment of Taman Posyandu Sukabumi , Wonosobo, Probolinggo, Bone	Almost Confident to Act
Jordan	2000/006	EVALUATION OF THE BETTER PARENTING PROGRAM	Confident to act
Kenya	2009/008	Evaluation of PHAST tool for the promotion hygiene Sanitation in the GOK/UNICEF Programme of cooperation	Almost Confident to Act
Kenya	2009/009	Supporting Sustainable water management and governance for the poor in drought and flood prone areas in Kenya	Almost Confident to Act
Kosovo	2009/002	Better Parenting Initiative Evaluation	Not Confident to Act
Kosovo	2009/003	Evaluation of the Education Project in Osterode	Almost Confident to Act
Lesotho	2009/002	The Evaluation of the Home Gardens	Almost Confident to Act
Madagascar	2009/003	Évaluation De La Mise En Oeuvre Et Proposition Des Orientations Pour Le Passage A L'echelle Du Programme Eka Y Compris L'informatisation De L'état Civil	Almost Confident to Act
Maldives	2009/006	Children and the 2004 Indian Ocean Tsunami: UNICEF's Response in Maldives – Country Synthesis Report	Confident to act
Mauritania	2009/007	Rapport d'Evaluation de l'Application de l'Ordonnance Portant Protection Pénale de l'Enfant en Mauritanie	Almost Confident to Act
Moldova	2008/010	Evaluation of Fight against Child Trafficking (FACT) Project	Almost Confident to Act
Moldova	2009/008	Evaluation of the Prevention of Mother to Child Transmission services in the Republic of Moldova	Almost Confident to Act
Mongolia	2009/002	Juvenile Justice Committee's Evaluation Report	Almost Confident to Act
Mozambique	2009/002	Preliminary Documentation and Evaluation of the Sanitation Component of the "One Million Initiative" Mozambique	Not Confident to Act
Namibia	2009/001	Evaluation of My Future is My Choice	Confident to act

Country	Sequence number	Title	Overall rating
Nepal	2009/015	Evaluation of the Partnership for Quality Education through Parental Participation	Confident to act
Nepal	2008/008	UNICEF Programme For The Reintegration Of Children Associated With Armed Forces And Armed Groups In Nepal - Evaluation Report: May 2008	Almost Confident to Act
Nepal	2009/009	Joint Evaluation of Nepal's Education for All 2004-2009 Sector Programme	Confident to act
Nepal	2009/011	Keeping Children in Focus (Strategic Evaluation of DACAW UNICEF Nepal)	Not Confident to Act
Nicaragua	2009/001	Evaluation of the actions to attend to the water sanitation and hygiene sector among populations affected by Hurricane Felix	Almost Confident to Act
Nicaragua	2009/002	Child Friendly Schools evaluation: Country Report for Nicaragua	Confident to act
Nigeria	2009/003	Child Friendly Schools Evaluation: Country Report for Nigeria	Almost Confident to Act
Occupied Palestinian Territory	2009/009	Palestinian Adolescents: agents of positive change- Towards an environment promoting peace and reconciliation	Almost Confident to Act
Occupied Palestinian Territory	2009/007	Evaluation of UNICEF-Supported Training Activities in Occupied Palestinian Territory (2006-2007)	Confident to act
Paraguay	2009/001	Programa Kits Escolares – Ministerio de Educación y Cultura	Not Confident to Act
Philippines	2009/001	Evaluation of the HIV Prevention Interventions for Most-at-Risk Adolescents	Almost Confident to Act
Philippines	2009/011	Child Friendly Schools Evaluation: Country Report for Philippines	Confident to act
Senegal	2009/001	Rapport d'Evaluation de la Mise en Place du Paquet de Services Intégrés dans les Ecoles Elémentaires des Régions de Ziguinchor, Kolda et Tambacounda	Confident to act
Serbia	2009/002	Evaluation of the Baby Friendly Hospital Initiative in Serbia for the period 1995-2008	Almost Confident to Act
Serbia and Montenegro	2009/001	Evaluation of Program – "School without Violence"	Not Confident to Act
South Africa	2009/001	Tree/UNICEF Kusaselihle Integrated Early Childhood Development Intervention (2004-2008)	Almost Confident to Act
South Africa	2009/010	Child Friendly Schools Evaluation: Country Report for South Africa	Confident to act
Sri Lanka	2009/001	Children and the 2004 Indian Ocean Tsunami: Evaluation of UNICEF's Response in Sri Lanka 2005-2008 – Country Synthesis Report	Confident to act
Sudan	2009/001	Improvement of the Health and Livelihood of Rural Communities in Southern Sudan and the Three Transitional Areas – European Commission UNICEF Project	Almost Confident to Act
Sudan	2008/007	Evaluation of UNICEF-GOS 2002-2006 Country Health and Nutrition Programme	Not Confident to Act
Swaziland	2009/001	Tinkhundla/constituencies fit for Children	Confident to act
Syria	2009/001	Program of Support to Syrian Education in areas affected by a large influx of Iraqui Refugee Children	Almost Confident to Act
Tanzania	2009/002	Evaluation of the Current Status and Future Utility of Cobet as a Strategic Intervention to Ensure Access to Quality Education for all Primary School-Ages Children in Tanzania	Not Confident to Act
Tanzania	2009/003	Evaluation Report of the Primary School Leaving Examination Conduct	Almost Confident to Act
Tanzania	2009/005	Evaluation of PMTCT Program in Refugee Camps in North Western Tanzania, 2003 - 2007	Not Confident to Act
Tanzania	2009/006	An Assessment of the Impact of Village/MTAA Resource Teams (BRTs) on the Activation of the Systems and Process in the Community within the Local Government Set-Up.	Not Confident to Act
Thailand	2009/004	Child Friendly Schools Evaluation: Country Report for Thailand	Confident to act
Thailand	2009/007	Children and the 2004 Indian Ocean Tsunami: UNICEF's Response in Thailand (2005-2008)	Very Confident to Act
Timor Leste	2009/012	Evaluation of the UNICEF Education Programme in Timor Leste	Very Confident to Act
Togo	2008/006	Rapport De L'évaluation Des Projets De Prévention Du VIH En	Confident to act

Country	Sequence number	Title	Overall rating
		Milieu Scolaire Dans Les Régions Maritime Et De La Kara	
Togo	2009/002	Evaluation de la couverture de la campagne nationale de distribution des moustiquaires imprégnées en 2008 et de l'impact des interventions de lutte contre le paludisme au Togo/Evaluation of the impact of anti malaria interventions, including 2008 national campaign to distribute ITNs	Almost Confident to Act
Uganda	2009/025	Final Review if UNICEF Supported (Hunter Foundation) Programmes for Children affected by Conflict in Kitgum, Northern Uganda	Confident to act
Uzbekistan	2009/002	Summative Evaluation of the Child Friendly Schools Project (2006-208)	Almost Confident to Act
Uzbekistan	2009/004	Evaluation of Family and Child Support Services Project	Confident to act
Uzbekistan	2009/005	Summative Evaluation of the Family Education Project for the period January 2005-July 2009	Very Confident to Act
Vietnam	2009/007	Evaluation of the Water Safety Model in IN THỪA THIÊN HUẾ	Almost Confident to Act
Vietnam	2009/012	Evaluation of Child-Friendly Primary Education	Almost Confident to Act
Vietnam	2009/017	Evaluation of the Pilot Project on Non-Custodial Measures, Reintegration and Support Services to Juveniles in Conflict with the Law in Hai Phone, Vietnam	Confident to act
Zambia	2009/001	Evaluation of the Community Based Orphan Support Programm3e & OVC Training: Chikankata Model	Not Confident to Act
Zambia	2009/003	Final Report on the Post Introduction Evaluation of the Pentavalent Vaccine in Zambia	Almost Confident to Act

Annex 4: Links to the review tool and to other online resources

1. UNICEF Global Evaluation Report Oversight System Methodology [INTRANET] (includes the Global Evaluation Report Oversight System Review Tool)

2. UNICEF Evaluation Policy [English]

3. UNICEF/UNEG Evaluation Report Standards [INTRANET]

4. Presentation on UNICEF Global Evaluation Report Oversight System [INTRANET]

5. UNICEF/UNEG Terms of Reference Standards [INTRANET]

6. Technical Note 1: Children Participating in Research and Monitoring and Evaluation (M&E) - Ethics and Your Responsibilities as a Manager

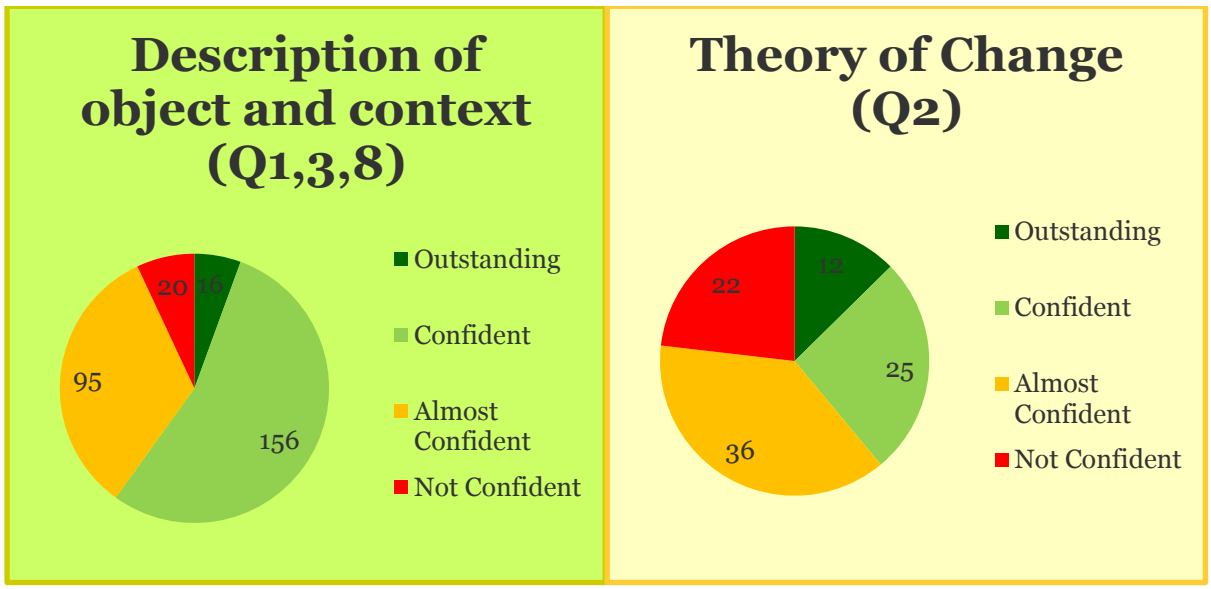
7. Technical Note 3: Writing a good Executive Summary

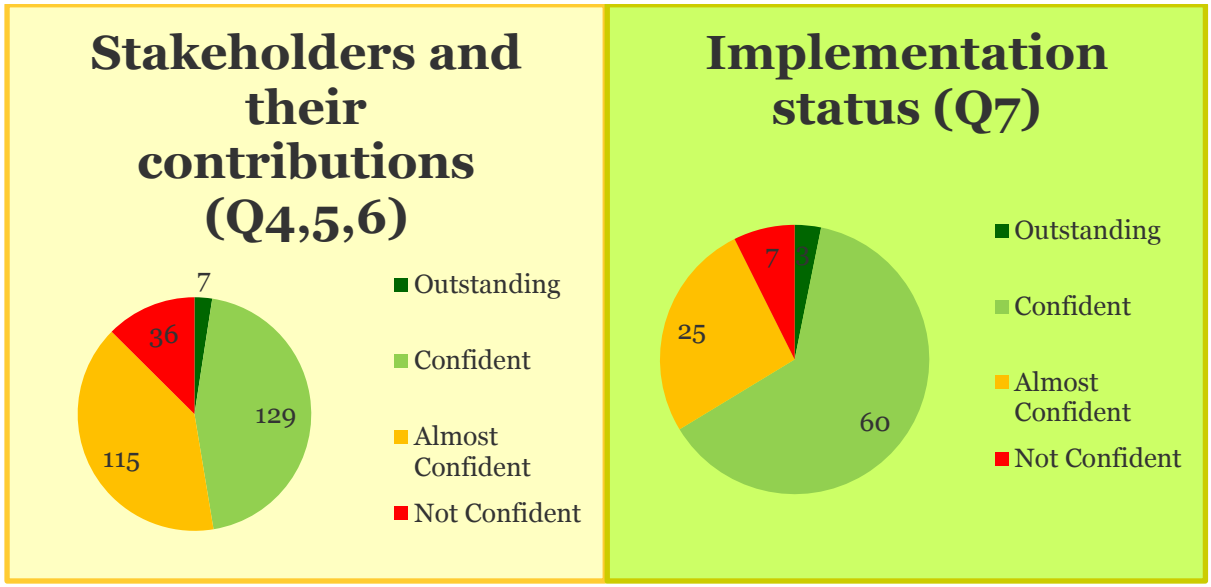
Annex 5: Performance dashboard of each cluster of questions

Clusters of questions relate to the main themes that commentary was given against and the UNEG/UNICEF Evaluation Standards refer to. **The question numbers that graphs are derived from are shown in parenthesis.** These graphs show relative ratings across the range of questions for each cluster and only for evaluation reports where a question was deemed applicable by the reviewer.

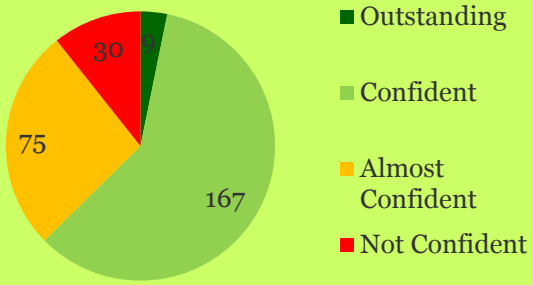
These graphs provide only one indication of performance and should be read in conjunction with the rest of this meta-evaluation report.

Background 'fill' colours are based on the median rating for that cluster. These provide a 'snap-shot' view of relative ratings and do not imply an overall level of satisfaction with that cluster.

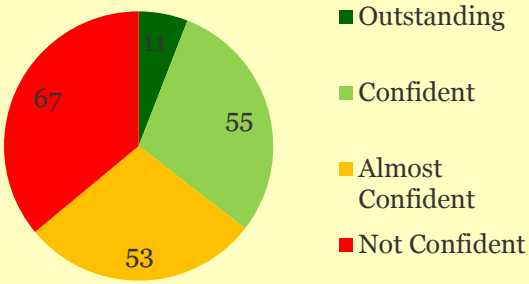




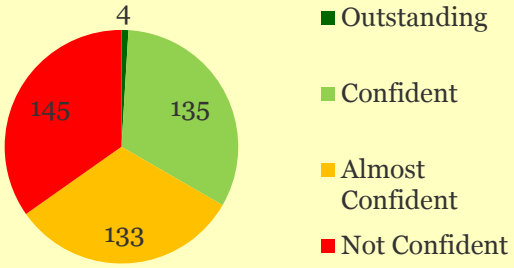
Purpose, objectives and scope (Q9,10,11)



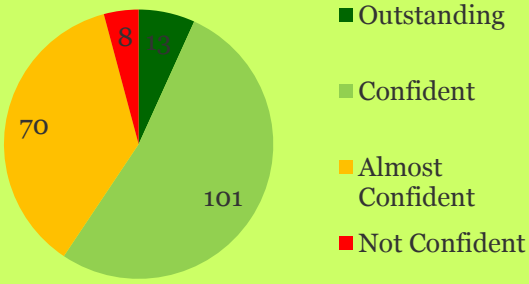
Evaluation Framework (Q12,13)



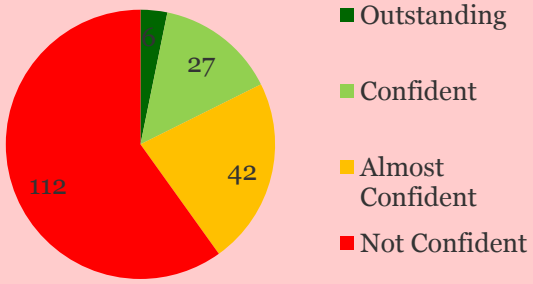
Human Rights, Gender and Equity (Q14,21,22,23,58)



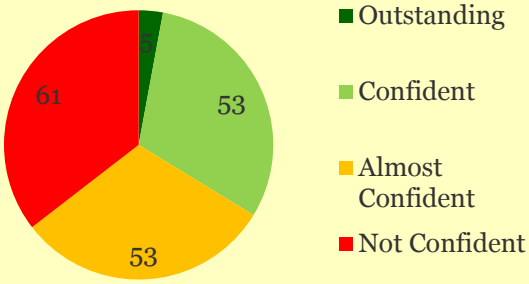
Data collection (Q15,16)



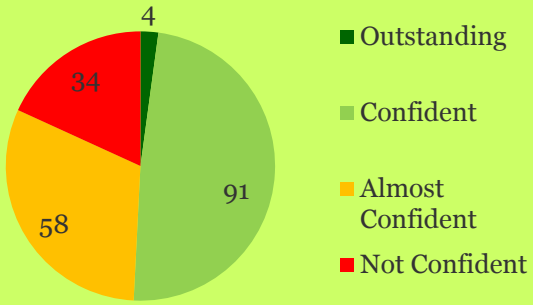
Ethics (Q17,18)



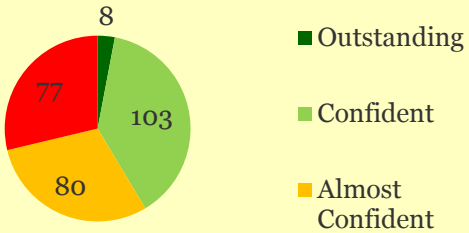
Results Based Management (Q19,20)



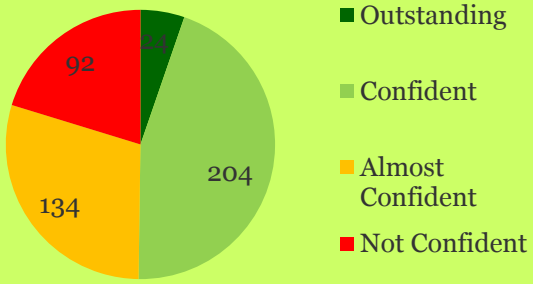
Stakeholder participation (Q24,25)



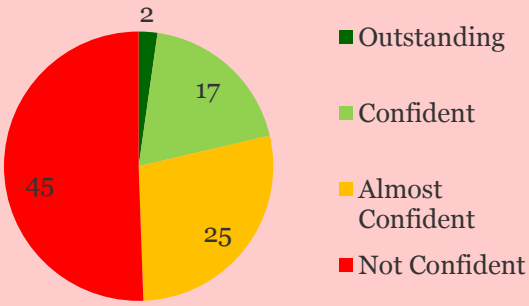
Methodological robustness, counterfactuals and limitations (Q26,27,28)



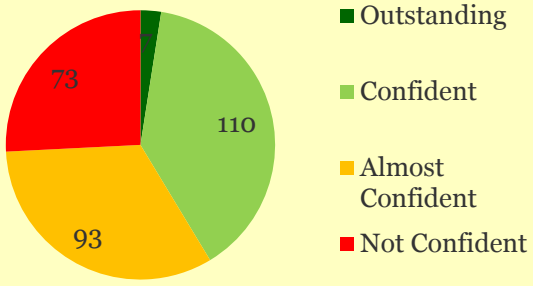
Completeness and logic of findings (Q29,30,31,32,33)



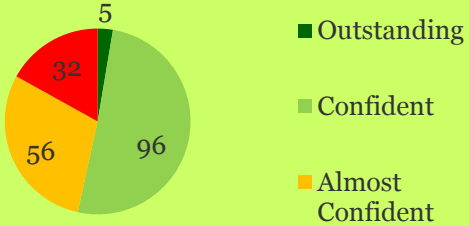
Cost analysis (Q34)



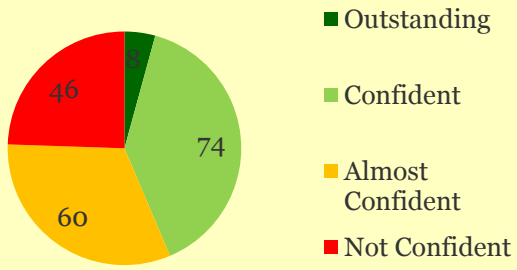
Contribution and causality (Q35,36,37)



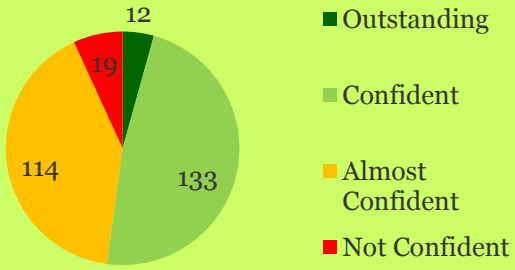
Strengths, weaknesses and future implications (Q38,39)



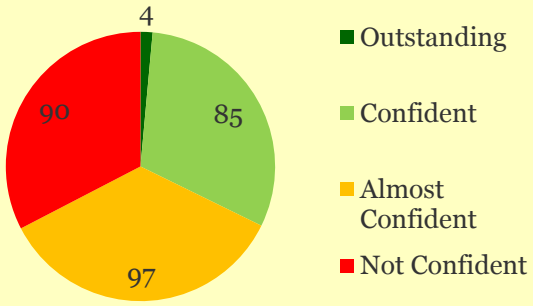
Completeness, level and insights of conclusions (Q40,41)



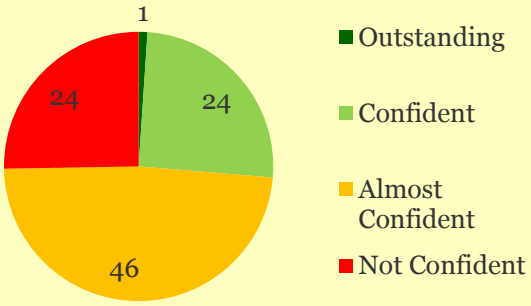
Relevance and clarity of recommendations (Q42,43,44)



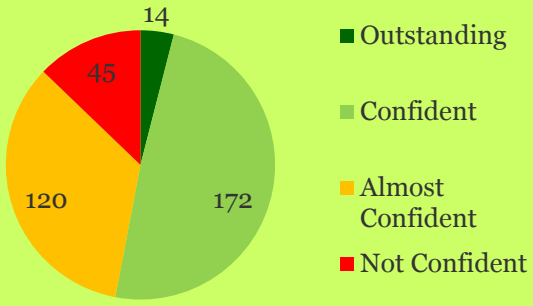
Usefulness of recommendations (Q45,46,47)



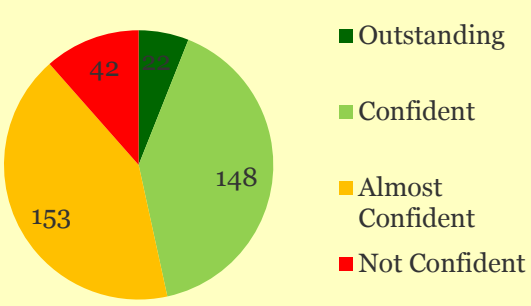
Appropriate lessons learned (Q48,49)



Executive Summary (Q51,52,53,54)



Style and presentation (Q50,55,56,57)



Annex 6: Experience of using the review tool and methodology

In addition to analysing information about the reviewed evaluation reports themselves, the meta-evaluation core team also collated and processed feedback from reviewers about the new iteration of the quality-assessment approach. This analysis generated four main findings relating to the viability of the qualitative approach, the handling of human rights and gender issues, the assessment of TORs, and the technical implementation of the tool.

The qualitative approach to evaluation report assessment is both viable and preferable to its predecessors

Although the methodology adopted for this meta-evaluation calls on established qualitative approaches, its adoption was experimental in this context. As a consequence, there was some uncertainty regarding the usefulness of qualitative analysis as compared to the quantified assessment previous undertaken, or whether issues of back-compatibility would be critical.

The overwhelming finding in this regard appears to be that not only is the new approach a viable alternative to using quantified scoring, but that it actually performs better as a review tool. Reviewers consistently reported being more comfortable with the new version of the tool; which, although taking slightly longer to complete, allowed them to draw out the real issues faced by each different evaluation report. By avoiding automated overall scoring, reviewers were able to treat each issue separately whilst still giving an accurate real-life appraisal of the overall confidence level.

Furthermore, although the review process generated a very large dataset this did create a barrier to useful analysis. Inclusion of report typology data and the harvesting of reviewer feedback made the identification and exploration of trends as viable, if not more viable, than the quantitative processing used in previous meta-evaluations. Where interesting issues were found, there was also the option to focus-in on the rich dataset in order to 'unpick' exactly what seemed to be the central issue.

All of this suggests that fully adopting this qualitative approach to evaluation report quality oversight is both viable and beneficial, and that there is no need to maintain direct links to the previous quantified assessment of the 22 UNICEF Evaluation Standards (the current design was intended to make it possible to 'reconstruct' the previous scoring if this had become necessary). Consequently, the opportunity presents itself for reviewing and further enhancing the review tool 'prompting' questions (the flexibility of the current methodology also makes this possible with compromising the comparability of overall ratings across versions).

Some suggestions from reviewers included the following *(the view of the core team is in parenthesis)*:

- Compressing both section C (methodology) and section D (findings and conclusions) to make them smaller by rationalising the questions into more prioritised aspects of the report *(this is recommended by the core team as a means of making space for disaggregated HRBAP and terms of reference questions)*;
- Reviewing the relationship between the executive feedback and constructive feedback 'boxes' and the purposes of these, with a view to preventing duplication (the constructive feedback column was felt to be one of the most valuable elements of the report) *(the core team recommend keeping these separate and will resend guidance that explains the different role of each box)*;
- Combining a number of questions, including Q12 with Q13, and Q48 with Q49;
- Combining the 'remarks boxes' for a number of questions, including Q38 with Q39 and Q40, Q51 with Q52-Q54, and Q15 with Q16;
- Investigating the option of having 'stop points' in the review (for example, if methodology is rated inadequate, then by definition the conclusions and recommendations cannot be relied upon) *(the core team do not yet believe that the review process is mature enough to deliver this successfully and that feedback should continue to be given on all sections)*; and

- Exploring the option of a 'light' version of the review format for small evaluations *(the core team believe that this would cause confusion in terms of classifying which evaluations are 'small')*.

The current review format is inadequate for fully exploring human rights, gender and equity issues; and needs to be linked to a mainstreaming strategy

As previously noted, the current approach of placing human rights questions at relevant points in the review tool was found to be unwieldy and not particularly coherent. Specifically, reviewers found that the human rights questions (Q14, Q21-Q23, Q58) did not work together as a 'whole' or 'fit' well within the tool. Furthermore, the analysis process found that it was challenging to distinguish gender and equity trends from these 'compound' questions.

An appropriate response to this issue needs to be more than 'clustering' the human rights, gender and equity questions under either an existing or new section of the review format. Rather, there appears to be a need to link the enquiry about these issues to a clearly articulated mainstreaming strategy for HRBAP in the evaluation function, which is manifested in coherent guidance for evaluators, explicitly linked-to in the terms of reference for evaluations, and an integral part of capacity building for the evaluation function. With such a framework in place, the review tool would be better positioned to adopt appropriately situated and coherent prompting questions to 'open-up' the 'black-box' of HRBAP issues in evaluation reports.

There is a direct nexus between terms of reference and the quality of evaluation reports, and a specific strategy needs to be deployed to address this issue

The adverse consequences of poor or missing terms of reference were continuously flagged as a matter of concern in reviews. This being the case, it was often suggested by reviewers that more value could be added through specifically assessing the terms of reference than the current format allows for. Data is currently collected on whether-or-not a terms of reference was included in the report, whether-or-not the report answers the terms of reference, and possible mitigating factors for the poor performance of a report. However, recurring concern about the quality of terms of reference, and the implications of this on evaluation reports, suggests that a case does exist for investigating options to specifically consider TOR quality within the review process.

The foundation of the technical solution to implementing the review tool is the right one and this can be further enhanced

The core meta-evaluation team experimented with a number of technical solutions to implementing the review tool, including web-based and Adobe PDF-based forms. After considering these alternatives, the final version of the tool was deployed using a Microsoft Excel template. This was found to have the following advantages:

- Familiarity of the format among reviewers;
- Easy access to the necessary software;
- Option of including explanatory information on additional worksheets within the template;
- Real-time population of the executive review format; and
- Ability to complete reviews 'off-line'.

A number of improvements are also possible based on the experience of this meta-evaluation that would reduce the need for post-completion processing. These are:

- The inclusion of an additional 'hidden' worksheet that converts the 'human-friendly' format of the existing tool into a 'database-friendly' format (i.e. data arranged in columns with consistent column headings) using cross-referencing;
- Pre-design of either an Excel 'master' workbook into which submitted reviews can be inserted; or the development of an Access Database that automatically imports all review data for post-processing;

- Pre-defined analytics requirements and formulas to enable testing prior to the deployment of the review templates; and
- Co-development and pre-testing of a Word-based executive feedback letter linked to either the master Excel workbook or the Access database using the mail-merge function.

In addition to the above options for enhancing the review format, the following two features were considered but could not be deployed without further development:

- *Automatic shading of ratings boxes* based on a conditional formatting rule is only compatible with Excel 2007 and later (.xlsx files) because of the need for at least four colour options. Many reviewers still use Excel 2003 (.xls files) and so this feature could not be adopted without encountering consistency problems.
- *Standard feedback statements for recurring issues* (such as including terms of reference) required both investigation into what these issues might be, and careful develop to ensure that the option to use a standard phrase did not encroach upon the provision of contextually-tuned feedback.

Annex 7: The review tool

UNICEF Global Evaluation Report Oversight System (GEROS) Review Template

Colour Coding	CC	Dark green	Green	Amber	Red	White
	Questions	Outstanding	Yes	Almost	No	Not Applicable
	Section & Overall Rating	Outstanding, best practice	Confident to act	Not quite confident to act	No confidence	

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

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

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

	Response				
Title of the Evaluation Report	**External summative evaluation of the Family Education Project for the period January 2005 - July 2009**				
Report sequence number	YYYY/123	**Date of Review**	DD/MM/YYYY	**Year of the Evaluation Report**	
Region				**Country(ies)**	
Type of Report				**TORs Present**	
Name of reviewer	**IOD PARC**				

Classification of Evaluation Report		Comments
Geographical *(Coverage of the programme being evaluated & generalizability of evaluation findings)*		
Management *(Managerial control and oversight of evaluation decisions)*		
Purpose *(Speaks to the overarching goal for conducting the evaluation; its raison d'etre)*		
Result *(Level of changes sought, as defined in RBM: refer to substantial use of highest level reached)*		
MTSP Correspondence *(Alignment with MTSP focus area priorities: (1) Young child survival and development; (2) Basic education and gender equality; (3) HIV/AIDS and children; (4) Child protection from violence, exploitation and abuse; and (5) Policy advocacy and partnerships for children's rights)*		
Level of Independence *(Implementation and control of the evaluation activities)*		
Timing / Stage		

SECTION A: OBJECT OF THE EVALUATION				
Question	**cc**	**Remarks**	**A/ Does the report present a clear & full description of the 'object' of the evaluation?** The report should describe the object of the evaluation including the results chain, meaning the 'theory of change' that underlies the programme being evaluated. This theory of change includes what the programme was meant to achieve and the pathway (chain of results) through which it was expected to achieve this. The context of key social, political, economic, demographic, and institutional factors that have a direct bearing on the object should be described. For example, the partner government's strategies and priorities, international, regional or country development goals, strategies and frameworks, the concerned agency's corporate goals & priorities, as appropriate.	**Constructive feedback for future reports** *Including how to address weaknesses and maintaining good practice*
1 Is the object of the evaluation well described? This needs to include a clear description of the interventions (project, programme, policies, otherwise) to be evaluated including how the designer thought that it would address the problem identified, implementing modalities, other parameters including costs, relative importance in the organization and (number of) people reached.				
2 Is the results chain or logic well articulated? The report should identify how the designers of the evaluated object thought that it would address the problem that they had identified. This can include a results chain or other logic models such as theory of change. It can include inputs, outputs and outcomes, it may also include impacts. The models need to be clearly described and explained.				
3 Is the context explained and related to the object that is to be evaluated? The context includes factors that have a direct bearing on the object of the evaluation: social, political, economic, demographic, institutional. These factors may include strategies, policies, goals, frameworks & priorities at the: international level; national Government level; individual agency level				
4 Are key stakeholders clearly identified? These include o implementing agency(ies) o development partners o rights holders o primary duty bearers o secondary duty bearers				
5 Are key stakeholders' contributions described? This can involve financial or other contributions and should be specific. If joint program also specify UNICEF contribution, but if basket funding question is not applicable				
6 Are UNICEF contributions described? This can involve financial or other contributions and should be specific				
7 Is the implementation status described? This includes the phase of implementation and significant changes that have happened to plans, strategies, performance frameworks, etc that have occurred - including the implications of these changes				
8 Does this illuminate findings? The context should ideally be linked to the findings so that it is clear how the wider situation may have influenced the outcomes observed.				
Executive Feedback on Section A Issues for this section relevant for feedback to senior management (positives & negatives), & justify rating. *Up to two sentences*				

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

12 Does the evaluation provide a relevant list of evaluation criteria that are explicitly justified as appropriate for the Purpose? It is imperative to make the basis of the value judgements used in the evaluation transparent if it is to be understood and convincing. UNEG evaluation standards refer to the OECD/DAC criteria, but other criteria can be used such as Human rights and humanitarian criteria and standards (e.g. SPHERE Standards) but this needs justification.. Not all OECD/DAC criteria are relevant to all evaluation objectives and scopes. The TOR may set the criteria to be used, but these should be (re)confirmed by the evaluator. Standard OECD DAC Criteria include: Relevance; Effectiveness; Efficiency; Sustainability; Impact Additional humanitarian criteria include; Coverage; Coordination; Coherence; Protection *(This is an extremely important question to UNICEF)*					
13 Does the evaluation explain why the evaluation criteria were chosen and/or any standard DAC evaluation criteria (above) rejected? The rationale for using each particular criterion and rejecting any standard OECD-DAC criteria (where they would be applicable) should be explained in the report.					
14 Did the evaluation design consider incorporation of the UN and UNICEF's commitment to a human rights-based approach to programming? This could be done in a variety of ways including: use of a rights-based framework, use of CRC, CEDAW and other rights related benchmarks, analysis of right holders and duty bearers and focus on aspects of equity, social exclusion and gender					
Executive Feedback on Section B Issues for this section relevant for feedback to senior management (positives & negatives), & justify rating. *Up to two sentences*					

SECTION C: EVALUATION METHODOLOGY, GENDER, HUMAN RIGHTS AND EQUITY				
Question	**cc**	**Remarks**	**C/ Is the methodology appropriate and sound?** The report should present a transparent description of the methodology applied to the evaluation that clearly explains how the evaluation was specifically designed to address the evaluation criteria, yield answers to the evaluation questions and achieve the evaluation purposes. The report should also present a sufficiently detailed description of methodology in which methodological choices are made explicit and justified and in which limitations of methodology applied are included. The report should give the elements to assess the appropriateness of the methodology. Methods as such are not 'good' or 'bad', they are only so in relation to what one tries to get to know as part of an evaluation. Thus this standard assesses the suitability of the methods selected for the specifics of the evaluation concerned, assessing if the methodology is suitable to the subject matter and the information collected are sufficient to meet the evaluation objectives.	**Constructive feedback for future reports** *Including how to address weaknesses and maintaining good practice*
15 Does the report specify data collection methods, analysis methods, sampling methods and benchmarks? This should include the rationale for selecting methods and their limitations based on commonly accepted best practice.				
16 Does the report specify data sources, the rationale for their selection, and their limitations? This should include a discussion of how the mix of data sources was used to obtain a diversity of perspectives, ensure accuracy & overcome data limits				
17 Are ethical issues and considerations described? The design of the evaluation should contemplate: How ethical the initial design of the programme was; The balance of costs and benefits to participants (including possible negative impact) in the programme and in the evaluation; The ethics of who is included and excluded in the evaluation and how this is done				
18 Does the report refer to ethical safeguards appropriate for the issues described? When the topic of an evaluation is contentious, there is a heightened need to protect those participating. These should be guided by the UNICEF Evaluation Office Technical Note and include: protection of confidentiality; protection of rights; protection of dignity and welfare of people (especially children); Informed consent; Feedback to participants; Mechanisms for shaping the behaviour of evaluators and data collectors				
19 Is the capability and robustness of the evaluated object's monitoring system adequately assessed? The evaluation should consider the details and overall functioning of the management system in relation to results: from the M&E system design, through individual tools, to the use of data in management decision making.				

20 Does the evaluation make appropriate use of the M&E framework of the evaluated object? In addition to articulating the logic model (results chain) used by the programme, the evaluation should make use of the object's logframe or other results framework to guide the assessment. The results framework indicates how the programme design team expected to assess effectiveness, and it forms the guiding structure for the management of implementation.						
21 Does the evaluation assess the extent to which the implementation of the evaluated object was monitored through human rights (inc. gender & child rights) frameworks? UNICEF commits to go beyond monitoring the achievement of desirable outcomes, and to ensure that these are achieved through morally acceptable processes. The evaluation should consider whether the programme was managed and adjusted according to human rights and gender monitoring of processes.						
22 Do the analytical framework, findings, conclusions, recommendations & lessons provide adequate information on human rights (inc. women & child rights) aspects? The inclusion of human rights and gender equality frameworks in the evaluation methodology should continue to cascade down the evaluation report and be obvious in the data analysis, findings, conclusions, any recommendations and any lessons learned.						
23 Is the methodology appropriate for analysing gender and human rights issues identified in the scope?If identified in the scope the methodology should be capable of assessing the level of: Identification of the human rights claims of rights-holders and the corresponding human rights obligations of duty-bearers, as well as the immediate underlying & structural causes of the non realisation of rights.; Capacity development of rights-holders to claim rights, and duty-bearers to fulfil obligations & aspects of social exclusion, disparities & equity.						
24 Are the levels and activities of stakeholder consultation described? This goes beyond just using stakeholders as sources of information and includes the degree of participation in the evaluation itself. The report should include the rationale for selecting this level of participation. Roles for participation might include: o Liaison o Technical advisory o Observer o Active decision making The reviewer should look for the soundness of the description and rationale for the degree of participation rather than the level of participation itself.						
25 Are the levels of participation appropriate for the task in hand? The breadth & degree of stakeholder participation feasible in evaluation activities will depend partly on the kind of participation achieved in the evaluated object. The reviewer should note here whether a higher degree of participation may have been feasible & preferable.						
26 Is there an attempt to construct a counterfactual? The counterfactual can be constructed in several ways which can be more or less rigorous. It can be done by contacting eligible beneficiaries that were not reached by the programme, or a theoretical counterfactual based on historical trends, or it can also be a comparison group.						
27 Can the methodology answer the evaluation questions in the context of the evaluation? The methodology should link back to the Purpose and be capable of providing answers to the evaluation questions.						
28 Are methodological limitations acceptable for the task in hand? Limitations must be specifically recognised and appropriate efforts taken to control bias. This includes the use of triangulation, and the use of robust data collection tools (interview protocols, observation tools etc). Bias limitations can be addressed in three main areas: Bias inherent in the sources of data; Bias introduced through the methods of data collection; Bias that colours the interpretation of findings						
Executive Feedback on Section C Issues for this section relevant for feedback to senior management (positives & negatives), & justify rating. *Up to two sentences*						

SECTION D: FINDINGS AND CONCLUSIONS					
Question	**cc**	**Remarks**	**D/ Are the findings and conclusions, clearly presented, relevant and based on evidence & sound analysis?** Findings should respond directly to the evaluation criteria and questions detailed in the scope and objectives section of the report. They should be based on evidence derived from data collection and analysis methods described in the methodology section of the report. Conclusions should present reasonable judgments based on findings and substantiated by evidence, providing insights pertinent to the object and purpose of the evaluation.		**Constructive feedback for future reports** *Including how to address weaknesses and maintaining good practice*
29 Are findings clearly presented and based on the objective use of the reported evidence? Findings regarding the inputs for the completion of activities or process achievements should be distinguished clearly from results. Findings on results should clearly distinguish outputs, outcomes and impacts (where appropriate). Findings must demonstrate full marshalling and objective use of the evidence generated by the evaluation data collection. Findings should also tell the 'whole story' of the evidence and avoid bias.	Yes				
30 Do the findings address all of the evaluation's stated criteria and questions? The findings should seek to systematically address all of the evaluation questions according to the evaluation framework articulated in the report.					
31 Do findings demonstrate the progression to results based on the evidence reported? There should be a logical chain developed by the findings, which shows the progression (or lack of) from implementation to results.					
32 Are gaps and limitations discussed? The data may be inadequate to answer all the evaluation questions as satisfactorily as intended, in this case the limitations should be clearly presented and discussed. Caveats should be included to guide the reader on how to interpret the findings. Any gaps in the programme or unintended effects should also be addressed.					
33 Are unexpected findings discussed? If the data reveals (or suggests) unusual or unexpected issues, these should be highlighted and discussed in terms of their implications.					
34 Is a cost analysis presented that is well grounded in the findings reported? Cost analysis is not always feasible or appropriate. If this is the case then the reasons should be explained. Otherwise the evaluation should use an appropriate scope and methodology of cost analysis to answer the following questions: o How programme costs compare to other similar programmes or standards o Most efficient way to get expected results o Cost implications of scaling up or down o Cost implications for replicating in a different context o Is the programme worth doing from a cost perspective o Costs and the sustainability of the programme.					
35 Does the evaluation make a fair and reasonable attempt to assign contribution for results to identified stakeholders? For results attributed to the programme, the result should be mapped as accurately as possible to the inputs of different stakeholders.					
36 Do conclusions take due account of the views of a diverse cross-section of stakeholders? As well as being logically derived from findings, conclusions should seek to represent the range of views encountered in the evaluation, and not simply reflect the bias of the individual evaluator. Carrying these diverse views through to the presentation of conclusions (considered here) is only possible if the methodology has gathered and analysed information from a broad range of stakeholders.					
37 Are causal reasons for accomplishments and failures identified as much as possible? These should be concise and usable. They should be based on the evidence and be theoretically robust. *(This is an extremely important question to UNICEF)*					
38 Are the future implications of continuing constraints discussed? The implications can be, for example, in terms of the cost of the programme, ability to deliver results, reputational risk, and breach of human rights obligations.					
39 Do the conclusions present both the strengths and weaknesses of the evaluated object? Conclusions should give a balanced view of both the stronger aspects and weaker aspects of the evaluated object with reference to the evaluation criteria and human rights based approach.					
40 Do the conclusions represent actual insights into important issues that add value to the findings? Conclusions should go beyond findings and identify important underlying problems and/or priority issues. Simple conclusions that are already well known do not					

add value and should be avoided.					
41 Are the conclusions pitched at a level that is relevant to the end users of the evaluation? Conclusions should speak to the evaluation participants, stakeholders and users. These may cover a wide range of groups and conclusions should thus be stated clearly and accessibly: adding value and understanding to the report (for example, some stakeholders may not understand the methodology or findings, but the conclusions should clarify what these findings mean to them in the context of the programme).					
Executive Feedback on Section D Issues for this section relevant for feedback to senior management (positives & negatives), & justify rating. *Up to two sentences*					
SECTION E: RECOMMENDATIONS AND LESSONS LEARNED					
Question	**cc**	**Remarks**	**E/ Are the recommendations and lessons learned relevant and actionable?** Recommendations should be relevant and actionable to the object and purpose of the evaluation, be supported by evidence and conclusions, and be developed with involvement of relevant stakeholders. Recommendations should clearly identify the target group for each recommendation, be clearly stated with priorities for action, be actionable and reflect an understanding of the commissioning organization and potential constraints to follow up.		**Constructive feedback for future reports** *Including how to address weaknesses and maintaining good practice*
42 Are the recommendations well-grounded in the evidence and conclusions reported? Recommendations should be logically based in findings and conclusions of the report.					
43 Are recommendations relevant to the object and the purpose of the evaluation? Recommendations should be relevant to the evaluated object					
44 Are recommendations clearly stated and prioritised? If the recommendations are few in number (up to 5) then this can also be considered to be prioritised. Recommendations that are over-specific or represent a long list of items are not of as much value to managers. Where there is a long list of recommendations, the most important should be ordered in priority.					
45 Does each recommendation clearly identify the target group for action? Recommendations should provide clear and relevant suggestions for action linked to the stakeholders who might put that recommendation into action. This ensures that the evaluators have a good understanding of the programme dynamics and that recommendations are realistic.					
46 Are the recommendations realistic in the context of the evaluation? This includes: o an understanding of the commissioning organisation o awareness of the implementation constraints o an understanding of the follow-up processes					
47 Does the report describe the process followed in developing the recommendations? The preparation of recommendations needs to suit the evaluation process. Participation by stakeholders in the development of recommendations is strongly encouraged to increase ownership and utility.					
48 Where presented, are lessons learned correctly identified? Lessons learned are contributions to general knowledge. They may refine or add to commonly accepted understanding, but should not be merely a repetition of common knowledge. Findings and conclusions specific to the evaluated object are not lessons learned.					
49 Where presented, are lessons learned generalised to indicate what wider relevance they may have? Correctly identified lessons learned should include an analysis of how they can be applied to contexts and situations outside of the evaluated object.					
Executive Feedback on Section E Issues for this section relevant for feedback to senior management (positives & negatives), & justify rating. *Up to two sentences*					

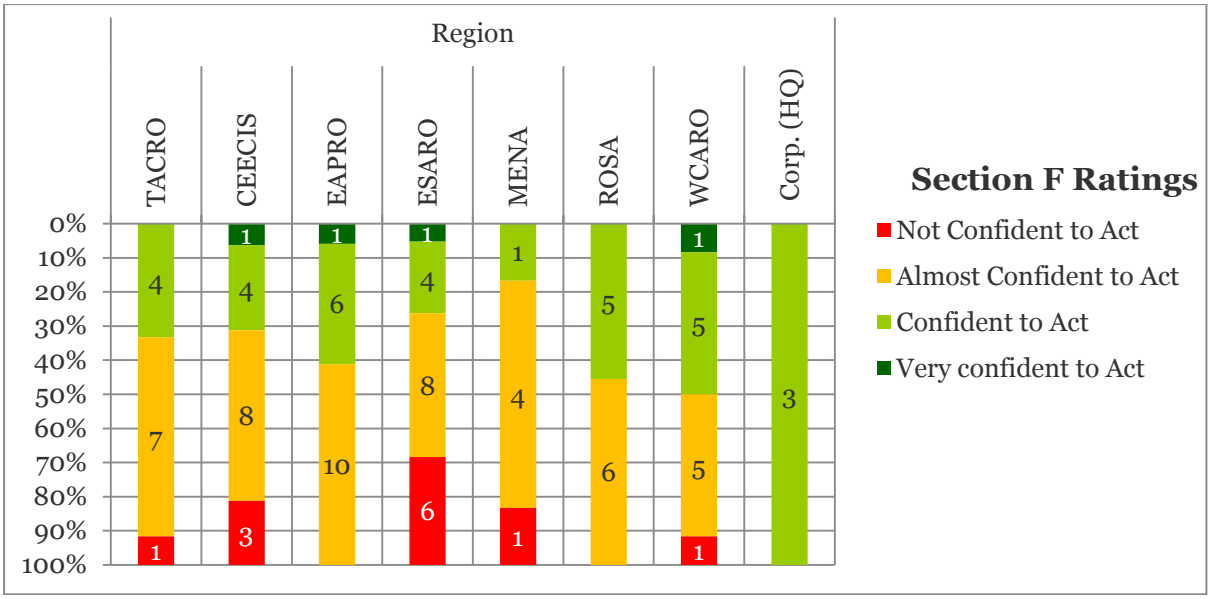
SECTION F: REPORT IS WELL STRUCTURED, LOGIC AND CLEAR					
Question	**cc**	**Remarks**	**F/ Overall, do all these elements come together in a well structured, logical, clear and complete report?** The report should be logically structured with clarity and coherence (e.g. background and objectives are presented before findings, and findings are presented before conclusions and recommendations). It should read well and be focused.		**Constructive feedback for future reports** *Including how to address weaknesses and maintaining good practice*
50. Do the opening pages contain all the basic elements? Basic elements include all of: Name of the evaluated object; Timeframe of the evaluation and date of the report; Locations of the evaluated object; Names and/or organisations of evaluators; Name of the organisation commissioning the evaluation; Table of contents including tables, graphs, figures and annex; List of acronyms					
51. Is an executive summary included as part of the report? If the answer is No, question 52 to 54 should be N/A					
52 Does the executive summary contain all the necessary elements? Necessary elements include all of: Overview of the evaluated object; Evaluation objectives and intended audience; Evaluation methodology; Most important findings and conclusions; Main recommendations					
53 Can the executive summary stand alone? It should not require reference to the rest of the report documents and should not introduce new information or arguments					
54 Can the executive summary inform decision making? It should be short (ideally 2-3 pages), and increase the utility for decision makers by highlight key priorities.					
55 Is the report logically structured? Context, purpose, methodology and findings logically structured. Findings would normally come before conclusions, recommendations & lessons learnt					
56 Do the annexes contain appropriate elements? Appropriate elements may include: ToRs; List of interviewees and site visits; List of documentary evidence; Details on methodology; Data collection instruments; Information about the evaluators; Copy of the evaluation matrix; Copy of the Results chain. Where they add value to the report					
57 Do the annexes increase the usefulness and credibility of the report?					
58 Is the style of the report human rights compliant? This includes: using human-rights language; gender-sensitive and child-sensitive writing; disaggregating data by gender, age and disability groups; disaggregating data by socially excluded groups					
Executive Feedback on Section F Issues for this section relevant for feedback to senior management (positives & negatives), & justify rating. *Up to two sentences*					

Additional Information	
Question	**Remarks**
i/ Does the evaluation successfully address the Terms of Reference? If the report does not include a TOR then a recommendation should be given to ensure that all evaluations include the TOR in the future. Some evaluations may be flawed because the TORs are inappropriate, too little time etc. Or, they may succeed despite inadequate TORs. This should be noted under vii in the next section	
ii/ Identify aspects of good practice of the evaluation In terms of evaluation	
iii/ Identify aspects of good practice of the evaluation In terms of programmatic, sector specific, thematic expertise	

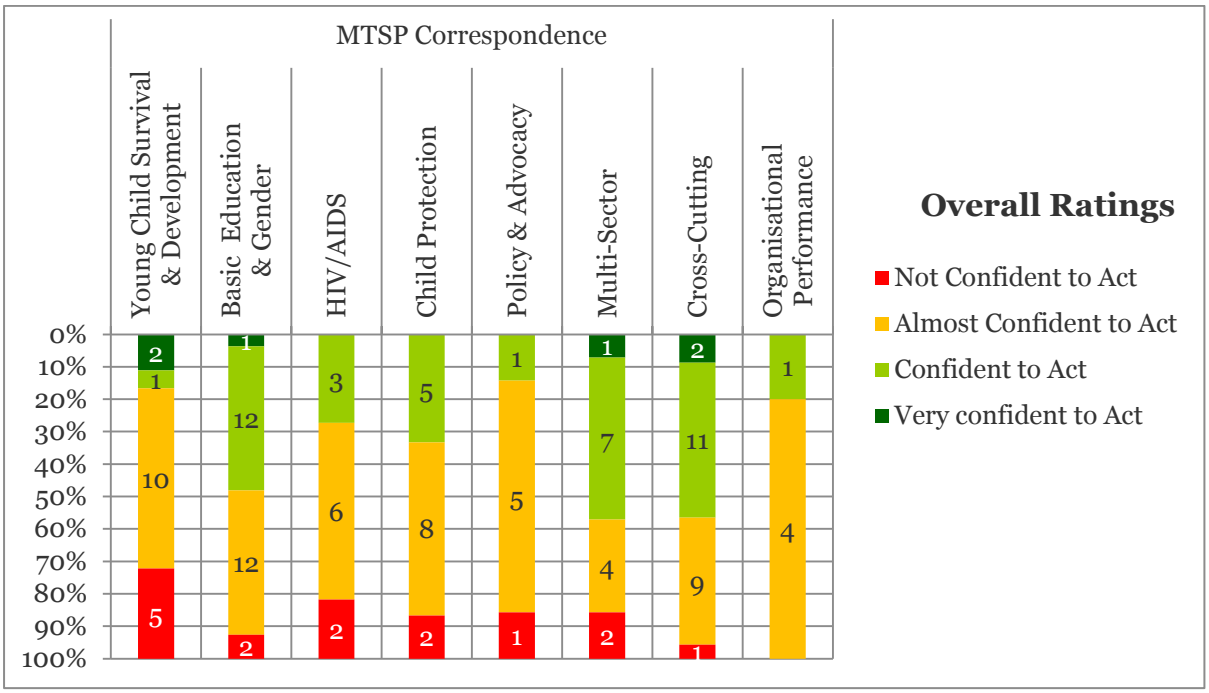
OVERALL RATING				
Question	**c c**	**Remarks**	**OVERALL RATING Informed by the answers above, apply the reasonable person test to answer the following question: Ω/ Is this a credible report that addresses the evaluation purpose and objectives based on evidence, and that can therefore be used with confidence?** This question should be considered from the perspective of UNICEF strategic management.	
i/ To what extent does each of the six sections of the evaluation provide sufficient credibility to give the reasonable person confidence to act? Taken on their own, could a reasonable person have confidence in each of the five core evaluation elements separately? It is particularly important to consider: o Is the report methodologically appropriate? o Is the evidence sufficient, robust and authoritative? o Do the analysis, findings, conclusions and recommendations hold together?				
ii/ To what extent do the six sections hold together in a logically consistent way that provides common threads throughout the report? The report should hold together not just as individually appropriately elements, but as a consistent and logical 'whole'.				
iii/ Are there any reasons of note that might explain the overall performance or particular aspects of this evaluation report? This is a chance to note mitigating factors and/or crucial issues apparent in the review of the report.				
Executive Feedback on Overall Rating Issues for this section relevant for feedback to senior management (positives & negatives), & justify rating. *Up to two sentences*				

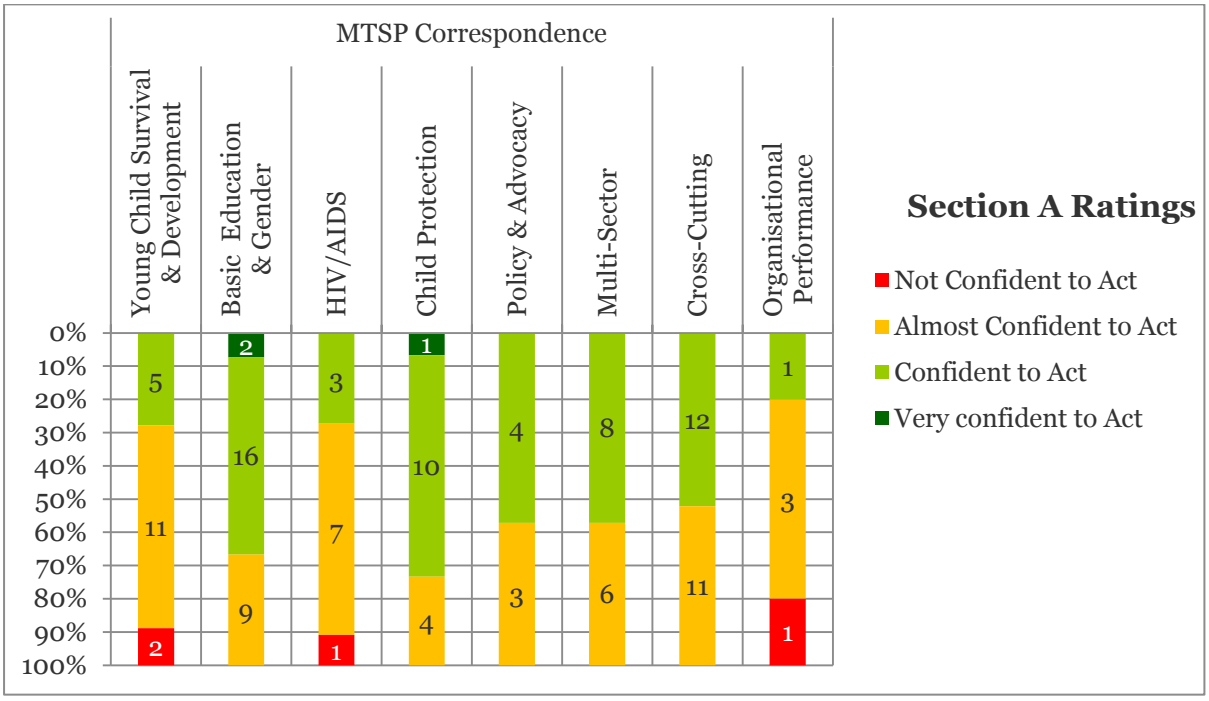
Annex 8: Regional breakdown of ratings by section and overall

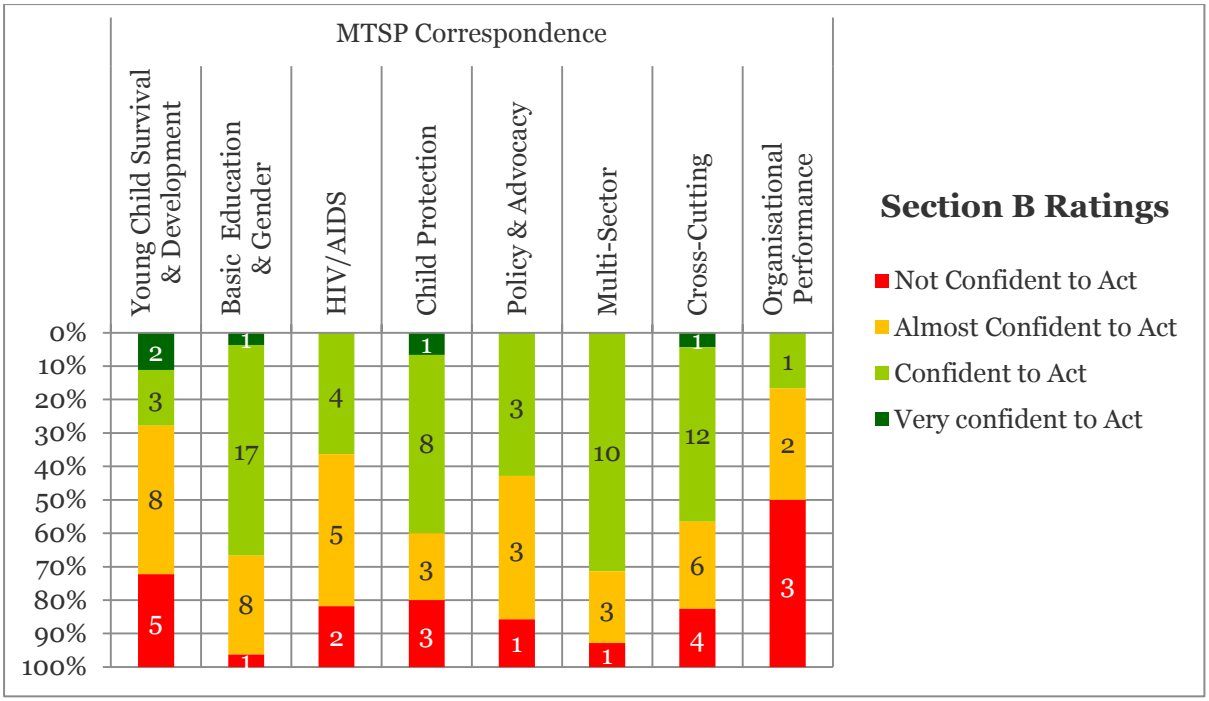

Region
TACRO
CEECIS
EAPRO
ESARO
MENA
ROSA
WCARO
Corp. (HQ)
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Section C Ratings
Not Confident to Act
Almost Confident to Act
Confident to Act
Very confident to Act
Section D Ratings
Not Confident to Act
Almost Confident to Act
Confident to Act
Very confident to Act
Section E Ratings
Not Confident to Act
Almost Confident to Act
Confident to Act
Very confident to Act

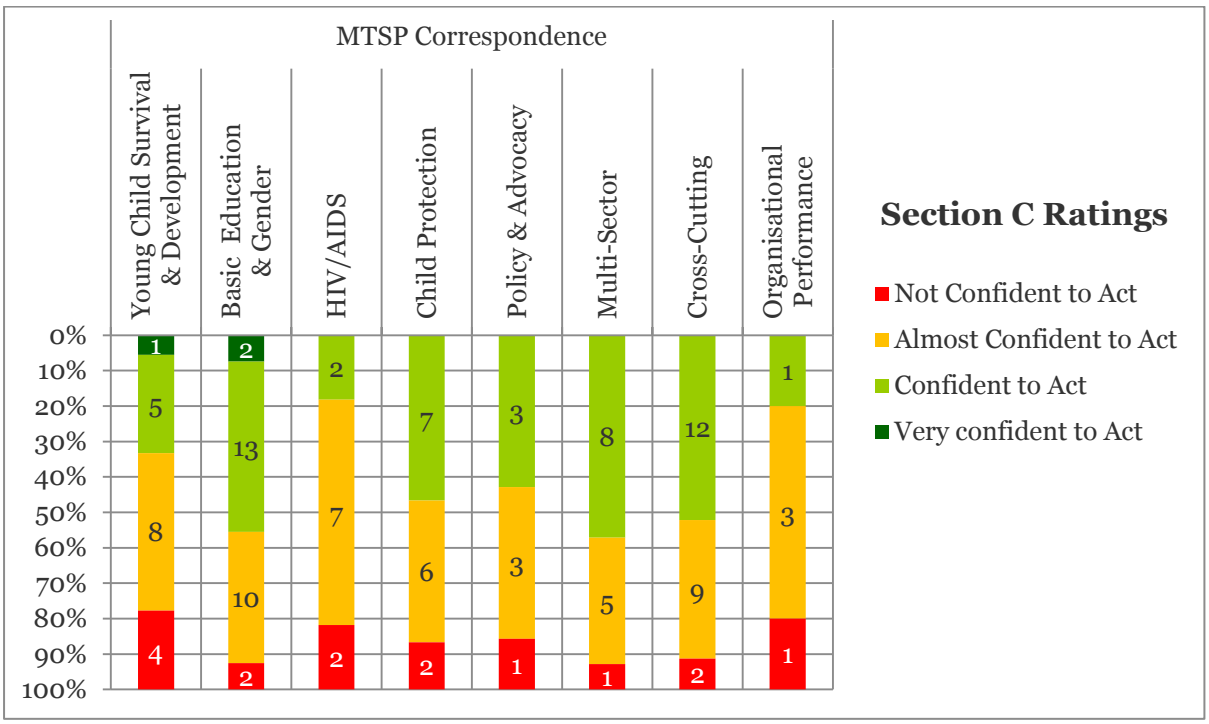

Region
TACRO
CEECIS
EAPRO
ESARO
MENA
ROSA
WCARO
Corp. (HQ)
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
4
7
1
1
4
8
3
1
6
10
1
4
8
6
1
1
4
1
5
6
1
5
5
1
3
Section F Ratings
Not Confident to Act
Almost Confident to Act
Confident to Act
Very confident to Act

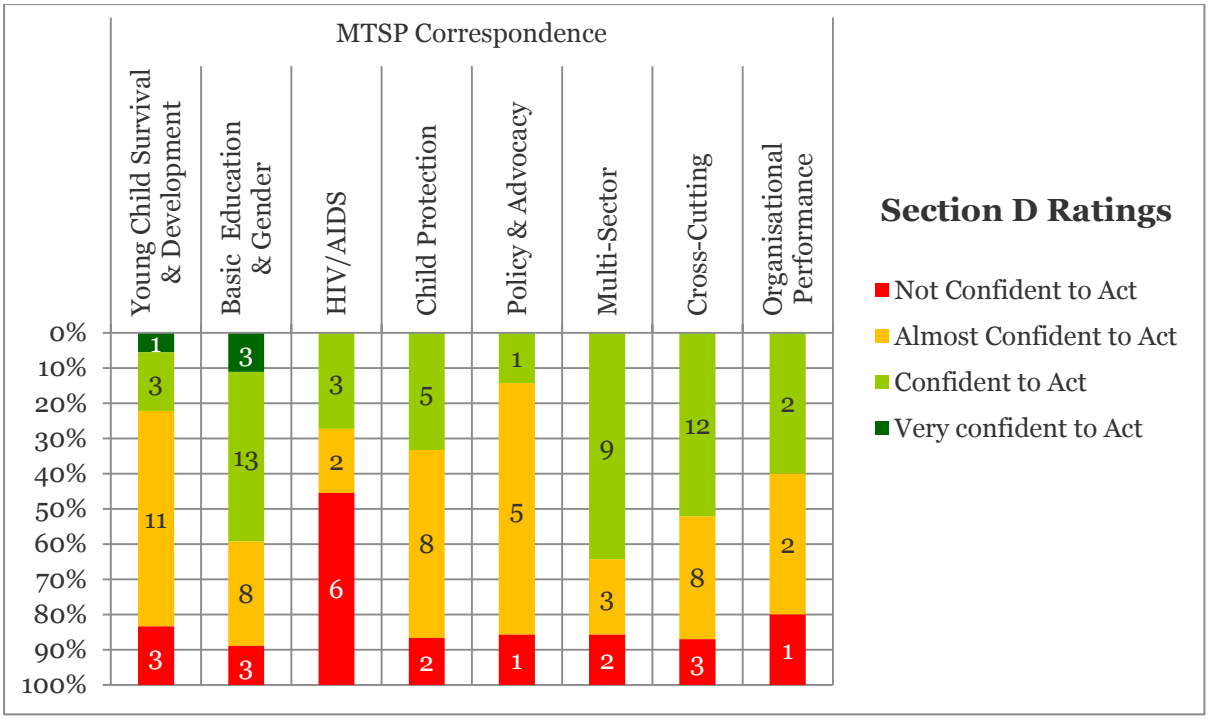
Annex 9: Breakdown of ratings according to MTSP-correspondence

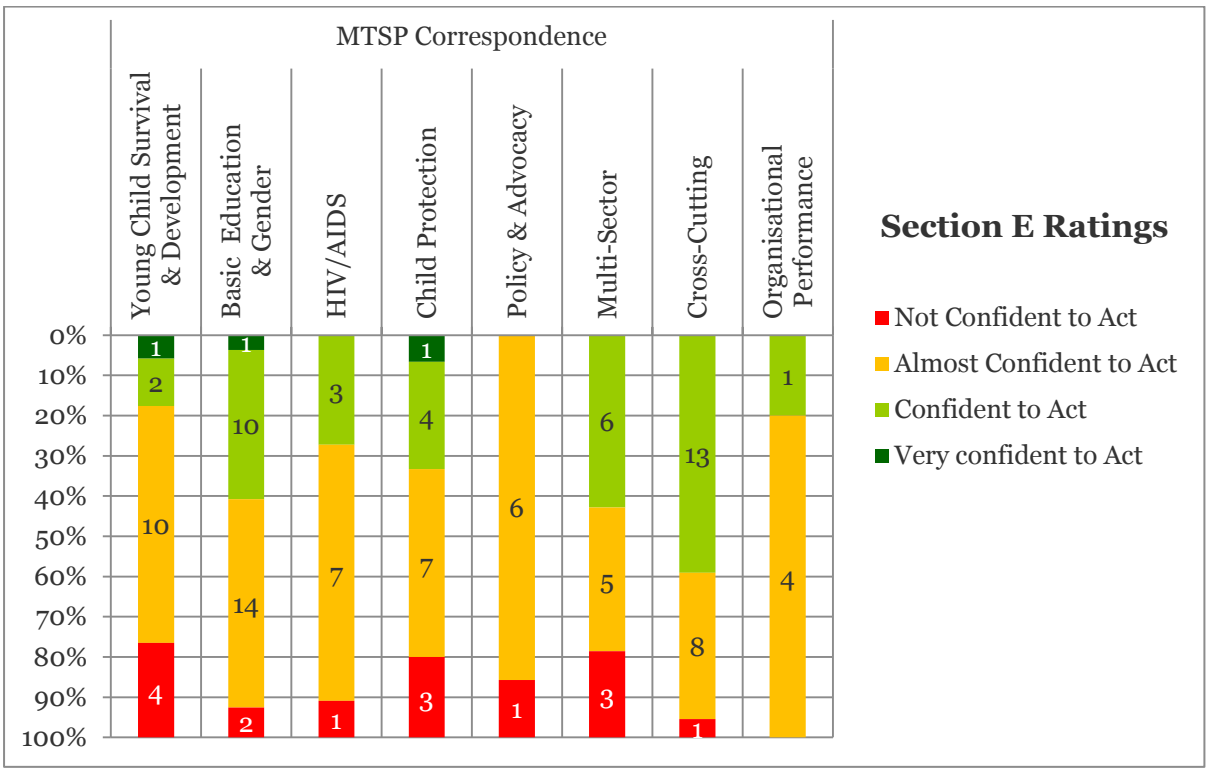


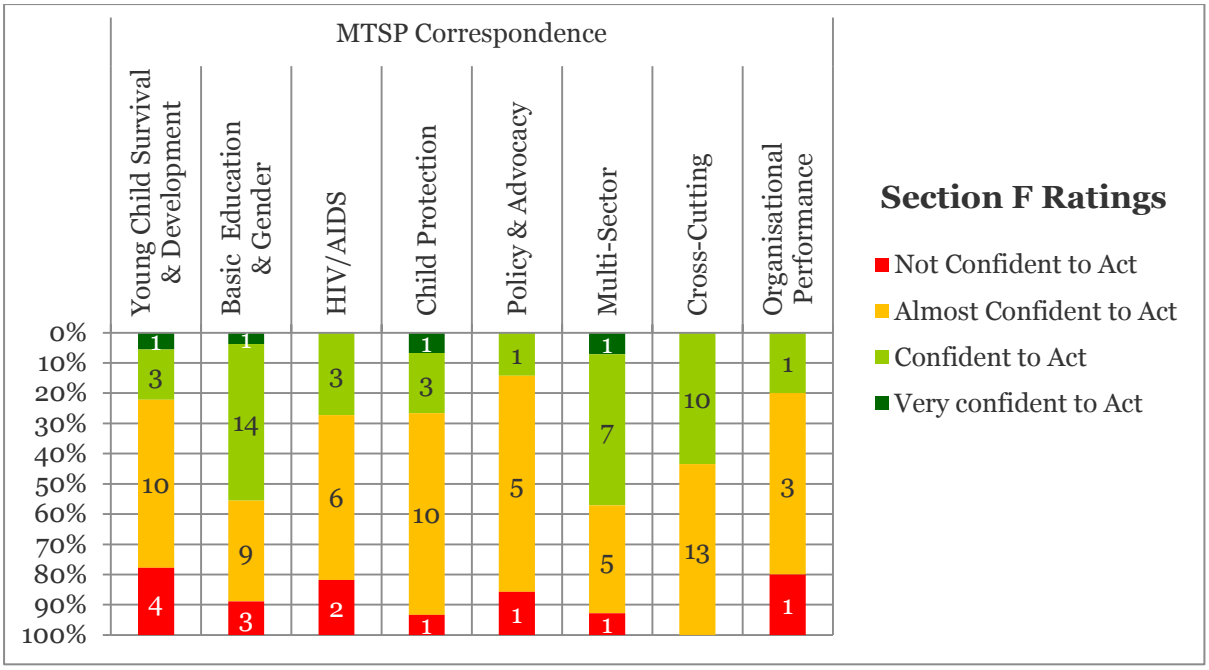



MTSP Correspondence
Young Child Survival & Development
Basic Education & Gender
HIV/AIDS
Child Protection
Policy & Advocacy
Multi-Sector
Cross-Cutting
Organisational Performance
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
2
3
8
5
1
17
8
1
4
5
2
1
8
3
3
3
3
1
10
3
1
1
12
6
4
1
2
3
Section B Ratings
Not Confident to Act
Almost Confident to Act
Confident to Act
Very confident to Act


MTSP Correspondence
Young Child Survival & Development
Basic Education & Gender
HIV/AIDS
Child Protection
Policy & Advocacy
Multi-Sector
Cross-Cutting
Organisational Performance
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
1
5
8
4
2
13
10
2
2
7
7
2
7
6
2
3
3
1
8
5
1
12
9
2
1
3
1
Section C Ratings
Not Confident to Act
Almost Confident to Act
Confident to Act
Very confident to Act


MTSP Correspondence
Young Child Survival & Development
Basic Education & Gender
HIV/AIDS
Child Protection
Policy & Advocacy
Multi-Sector
Cross-Cutting
Organisational Performance
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
1
3
11
3
3
13
8
3
3
2
6
5
8
2
1
5
5
1
9
3
2
12
8
3
2
2
2
1
Section D Ratings
Not Confident to Act
Almost Confident to Act
Confident to Act
Very confident to Act


MTSP Correspondence
Young Child Survival & Development
Basic Education & Gender
HIV/AIDS
Child Protection
Policy & Advocacy
Multi-Sector
Cross-Cutting
Organisational Performance
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
1
2
10
4
1
10
14
2
3
7
1
1
4
7
3
1
6
1
6
5
3
13
8
1
1
4
Section E Ratings
Not Confident to Act
Almost Confident to Act
Confident to Act
Very confident to Act


MTSP Correspondence
Young Child Survival & Development
Basic Education & Gender
HIV/AIDS
Child Protection
Policy & Advocacy
Multi-Sector
Cross-Cutting
Organisational Performance
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
1
3
10
4
1
14
9
3
3
6
2
1
3
10
1
1
5
1
1
7
5
1
10
13
1
3
1
Section F Ratings
Not Confident to Act
Almost Confident to Act
Confident to Act
Very confident to Act