OCB META EVALUATION OF

SEU-MANAGED EVALUATIONS 2017-2021

DECEMBER 2022

This publication was produced at the request of MSF-OCB under the management of the Stockholm Evaluation Unit.

It was prepared independently by Michael Harnar, Zach Tilton and Tian K. Ford.

DISCLAIMER

The authors' views expressed in this publication do not necessarily reflect the views of Médecins sans Frontières and the Stockholm Evaluation Unit.

TABLE OF CONTENTS

List of Figures/Tables 2

Acronyms 3

Definitions 4

EXECUTIVE SUMMARY 6

INTRODUCTION 7

FINDINGS 16

MEQ1 (Defining Quality): What are the most important, relevant, and useful meta-evaluation criteria for MSF in determining the quality and value of past and future evaluations? 17

MEQ2 (Assessing Quality): To what extent do the products and performances of OCB finalized evaluations from 2017-2021 meet generally accepted evaluation quality criteria and standards? 19

MEQ3: With these meta-evaluative conclusions, to what extent do these completed evaluations provide value to OCB? 45

MEQ4: What factors within MSF organizational sphere of influence mediate the quality of evaluation processes and products and how can MSF use this information to ensure high quality and value evaluations? 52

CONCLUSIONS 60

RECOMMENDATIONS 62

ANNEXES 64

LIST OF FIGURES/TABLES

Figure 1: METAE meta-evaluation framework pyramid 12
Figure 2. Evaluation Utility, Use Types and Use Outcomes 46
Figure 3. Budget and Utility z-Scores 51

Table 1. Example criterion, sub-criterion, and indicators 13
Table 2. Values about evaluation quality from SEU evaluation policy 18
Table 3. Detailed composition of METAE meta-evaluation framework..... 19
Table 4. PrgES Criteria Ranks, Scores, and Ratings..... 20
Table 5. ALNAP Proforma Domain Ratings, Scores, and Ratings..... 25
Table 6. UNEG Norms Definitions and Ratings..... 29
Table 7. UNEG Standards and Sub-standards Ratings 35
Table 8. EMQF Ratings for SEU Evaluation Portfolio and System 41
Table 9. Scores and Ratings of Recorded PrgES and ALNAP Indicators to Modified EMQF Domains 44
Table 10. Quality Gap list Represented by PrgES Indicators with Scores of 25% or Less 53
Table 11. Six-step evaluation scores 55

ACRONYMS

MSF	Médecins Sans Frontières
OCB	Operational Centre Brussels
SEU	Stockholm Evaluation Unit
SC	(SEU) Steering Committee
PrgES	Program Evaluation Standards
ALNAP	Active Learning Network for Accountability and Performance
EMQF	SEU Evaluation Manifesto Quality Framework
CP	Checkpoint
TOR	Terms of Reference
IR	Inception Report
FR	Final Report
MR	Management Response
MEQ	Meta-evaluation Question
A8CP1	An example of a reference point within the PrgES checklist. The first letter, “A”, refers to the PrgES criteria, and “8” references the relevant sub-criteria. “CP1” refers to the first checkpoint of the sub-criteria.
AEA	American Evaluation Association
EHA	Evaluating Humanitarian Action

DEFINITIONS

Evaluation – the product or process of determining the merit, worth, or significance of something.
Meta-evaluation – the evaluation of evaluations.
Values – values are beliefs about what is important held by individuals and groups. They inform value judgments and are the result of social, cultural, and psychological factors.
Quality – the intrinsic merit of anything. Claims of quality result from judgments that compare actual performance with absolute dimensions of merit.
Value – the extrinsic worth of anything. Claims of value result from judgments that compare actual performance with relative standards or comparators.
Principle – a value expressed as a verb. Prescriptive action statements of what ought to be done ideally such as "speak out" or "obtain informed consent".
Criteria – dimensions of quality or merit. Criterion is the singular form.
Sub-criteria – facets of dimensions or criteria of merit.
Quality Domain – these are synonymous with criteria in this study.
Quality Element – these are synonymous with sub-criteria in this study.
Norm – a standard or pattern that is typical or expected of a group. These are synonymous with criteria in this study.
Standard – Generally a dimension of quality established by some authority as a rule for the measure of quantity, weight, extent, value, or quality. Two quality frameworks in this study (PrgES and UNEG) use this term to refer to criteria and sub-criteria of merit. For clarity the evaluation team uses this term to refer to the degrees or levels of merit such as *Poor, Fair,* and *Good.*
Indicator – an observation that describes the state or level of something.
Checkpoint – a statement that implies the presence or absence of something in an evaluation that is either "met" nor "not met", respectively. For this study the terms checkpoint and indicator are synonymous.
Transversal – a descriptor used to classify evaluations or analyses that focus on areas where two or more projects have the same activities and comparison is possible. The area of focus can include Models of Care (e.g. Victim of Torture - VOT, Non-Communicable Diseases - NCD) or thematic areas (e.g. migration).
Evaluation Quality Framework – a collection of performance criteria, sub-criteria, and indicators of merit that describe the ideal notion of a particular object of evaluation.
Evaluation Case – each evaluation under review in this meta-evaluation is considered a case.
Evaluation Portfolio – referred to as the evaluation dossier within SEU, this is a shorthand for a collection of ongoing or past evaluations. In this meta-evaluation report, portfolio refers to the 31 completed and reviewed primary evaluations over the past 5 years.
Evaluation System – a set of institutional assets such as policies, guidelines, procedures, human and financial resources used to manage or conduct evaluations.
Evaluation Culture – the attitudes toward and behaviors associated with evaluation within an organization.
Evaluation Policy – a stated organizational position that articulates the ideal conduct of evaluation. This can refer to a specific position or a collection of coherent positions.
Formative evaluation – improvement-oriented evaluation that often happens before or during the development or execution of some performance or product.

Summative evaluation – judgment-oriented evaluation that often happens at some key decision point related to significant commitments, typically after the conclusion of some performance or product delivery

Evaluation use – broad term that implies any changes to individual and or organizational attitudes and behaviors as a result of evaluation processes and products.

Process use – intentional and unintentional changes that occur to individual and organizational attitudes and behaviors as a result of the process of evaluation.

Findings use – intentional and unintentional changes that occur to individual and organizational attitudes and behaviors as a result of the application of evaluation findings and recommendations.

Conceptual use – enhanced understanding of the nature of evaluation or of the object of evaluation in evaluation participants due to evaluation processes or products.

Instrumental use – changes of behavior in evaluation participants due to evaluation processes or products.

Evaluation capacity development use – increased capacity to conduct, manage, or use evaluations in evaluation participants due to evaluation processes or products.

Legitimative use – conducting an evaluation to confirm or support decisions already made in advance of evaluation processes or findings.

Symbolic use – commissioning or conducting an evaluation for political show, public relations, to placate individuals or organizations, or delay action.

Evaluation use outcomes – for this study, use outcomes refers to any effect or consequence resulting from changes in individual or organizational attitudes or behaviors due to evaluations—what results from the process or findings use of evaluations.

>-<

EXECUTIVE SUMMARY

The following report presents findings, conclusions, and recommendations from an external independent meta-evaluation study conducted from May to October 2022. The goal of this meta-evaluation was to assess the quality and value of past evaluations and produce useful processes and products to "maximiz[e] the strategic value of requesting, conducting and making use of [future] evaluations" at Operational Centre Brussels.

Object of Meta-evaluation

The primary object of this meta-evaluation was the portfolio of evaluations managed by the SEU from 2017-2021; the secondary object of this meta-evaluation was the evaluation system at OCB.

Meta-evaluation Objectives

1. Assess quality and value of past evaluations.
2. Establish understanding of evaluation quality at OCB.
3. Determine organizational factors of evaluation quality.

Meta-evaluation Questions

1. How does the SEU and OCB **define** evaluation quality?
2. What was the **quality** of past evaluation performance?
3. What was the **value** of past evaluation performance?
4. What **factors** determined the quality of past evaluation performance?

Methods

The design of this meta-evaluation is an external and independent ex-post portfolio meta-evaluation. The meta-evaluation framework is a hybrid of the Program Evaluation Standards, the ALNAP Proforma, the UNEG Norms and Standards, and the SEU Evaluation Manifesto Quality Framework. Methods used for data collection were document review, key informant interviews, focus group discussion, online survey questionnaires, and participant observation. Checklist and numerical weight and sum methodologies were used for analysis and synthesis.

Findings

PrgES Portfolio Rating and Score: **GOOD (60%)**
ALNAP Portfolio Rating and Score: **VERY GOOD (80%)**
UNEG Rating: **GOOD**
EMQF Rating: **VERY GOOD**

Conclusions

Definitions of evaluation quality at OCB are emerging and defensible. Evaluation quality at OCB is **GOOD** to **VERY GOOD**. Evaluation use at OCB is **GOOD**. Outcomes of evaluation use are **FAIR**. The full extent of evaluation use and outcomes is still unknown. The OCB is receiving **GOOD** value from the evaluation function. The evaluation system at OCB is well functioning and healthy. Findings and recommendations from this meta-evaluation provide OCB a roadmap for sustaining and improving quality and value.

Recommendations

Recommendation 1: Strengthen the **Evaluability Assessment** Function
Recommendation 2: Re-invest in Documenting Evaluation **Use and Influence**
Recommendation 3: Demand Stronger **Evaluative Logic, Reasoning, and Valuing**
Recommendation 4: Formalize the **Internal Meta-evaluation Function**
Recommendation 5: Adopt **Transformative Evaluation** Policies

INTRODUCTION

Quality programs, program evaluations, and evaluative functions at Médecins Sans Frontières (MSF) are warranted by a host of institutional, operational centre-, and unit-specific commitments. The foundation of these commitments to quality evaluation and evaluation culture are found implicitly in movement-wide policy and operational documents such as The Charter (1971), Chantilly Principles (1997), and La Mancha Agreement (2006). For the Operational Centre Brussels (OCB) they are further legible in documents such as the Policy on Quality Improvement (2016), OCB Quality of Care- Policies and Tools, and the MSF-OCB Quality Framework. Finally, explicit commitments to quality evaluation at the OCB are codified across multiple evaluation policy and guideline documents published by The Stockholm Evaluation Unit (SEU).[1] These principles, policies, and procedures form the basis of quality evaluation culture within the movement and specifically OCB, driven by the SEU. This culture of evaluation is best expressed by the La Mancha Agreement that states MSF "is committed to the impact and effectiveness of its work so that good work can be multiplied and abandon ineffective practice."

The OCB is committed to developing organizational evaluation culture and capacities to enable operations teams to learn and improve operational and medical interventions. The OCB recognizes that the effectiveness of programs is based on the effectiveness of the design, monitoring, and evaluation of those programs and requires the evaluation of evaluations—meta-evaluation—to responsibly fulfill commitments made by those in the movement as reflected in the OCB Strategic Directions. The following report presents findings, conclusions, and recommendations from an external independent meta-evaluation study conducted from May to October 2022. The goal of this meta-evaluation is to assess the quality and value of past evaluations and produce useful processes and products to "maximiz[e] the strategic value of requesting, conducting and making use of [future] evaluations" at the Operational Centre Brussels.[2]

META-EVALUATION OBJECT BACKGROUND AND DESCRIPTION

EVALUATION SYSTEM

The Operational Centre Brussels (OCB) is one of six Operational Centres within the Médecins Sans Frontières International movement of 26 different member associations and legal entities operating under the same name of MSF. OCB includes nine MSF associations and six branch offices across Europe, Middle East, Africa, Latin America, and Asia. The Centre has a Board of Directors and General Direction team of executives and support units that oversee eight departments, including the Operations Department, which uses eight regional cells, an emergency pool, and operations support. The Stockholm Evaluation Unit (SEU), based in Sweden, is one of three MSF units tasked to manage and guide evaluations of MSF's operational projects. The SEU is housed within the Operations Support Team within the Operations Department of the OCB and consists of one head of unit, three evaluation managers, and one coordinator. The functional management of the SEU is overseen by the SEU Steering Committee (SC) that ensures SEU independence and accountability for SEU quality of work

[1] These include: The SEU Evaluation Framework; The SEU Evaluation Manifesto; The SEU Ethical Guidelines; The Six-Step Process; Roles and Responsibilities; the SEU/OCB Strategy and Governance; SEU Steering Committee ToR.
[2] SEU Steering Committee ToR.

and progress. The SC consists of a chair and representatives from OCB Board, Director General Office, representatives from the operations and medical departments, and two representatives of MSF Sweden.[3]

The SEU has a distinctive and generally accepted point of view about the nature and purpose of evaluation. For the SEU, evaluation is "a systematic process to judge merit, worth or significance by combining evidence and values. Simply stated, evaluation is for the sake of making a judgment: what was valuable, why was it good, how was it successful?"[4] This explanation is likely helpful for the unit and OCB which has other systematic disciplined inquiry modes for operation support such as capitalizations, lessons learned studies, research studies, and audits. These non-evaluative studies are not managed by the SEU but by other projects, departments, cells, and missions.

The SEU evaluates to drive quality and improve operational decision making to better serve the intended beneficiaries of OCB with lifesaving medical humanitarian interventions. This is informed by the *2020-23 OCB Strategic Orientations* document, which commits to "a culture of evaluation to give the field teams the opportunity to learn from [their] practices and to constantly improve the quality and pertinence of operational/medical interventions." Further, the SEU is in a unique position as an institutional evaluation unit in that MSF does not receive funding from bilateral donors. This means evaluations managed by SEU are not mandated by external conditionalities or funding requirements, but by a "commitment to transparency, underscoring our accountability towards external stakeholders including patients and the community, partnering organizations, as well as donors."[5]

OCB recently adopted a default evaluation policy that "all projects should be evaluated during their lifespan unless there is a well justified reason not to do so."[6] Although OCB was implementing an estimated 70+ ongoing projects at various stages in early 2022, the SEU has set a guideline to manage roughly 15-20 evaluations per year.[7] Though the object of evaluations vary, OCB prioritizes evaluating its medical humanitarian operations. Evaluations can and do happen at any point of a project lifecycle at OCB with a preference given to formative or mid-term evaluations. There are five main evaluation types the SEU manages, which include: *mid-term evaluations*, typically for multi-year projects; *end-of-project evaluations* that are goal- and outcome-oriented and intended to support handover or closure strategy; *pilot and innovation project evaluations* focused on context-specific and transferable knowledge to guide future decisions for project development; emergency projects that are mostly real-time evaluations intended to check assumptions and inform ongoing strategy; and *transversal thematic or model of care evaluations* that are a type of cluster or portfolio evaluation looking across multiple projects within the same or similar family of medical humanitarian interventions. These transversal evaluations are geared toward cross-context knowledge translation and policy development. All evaluations consist of a six-step process that is managed by the SEU.[8] The combination of this institutional context, organizational structure, policies, procedures, and human and financial resources constitute what can be termed an evaluation system.

The state of evaluation quality is a central consideration for this evaluation system and has been an operational priority for the SEU from its inception as evidenced by various policies and procedures for

[3] Ibid.
[4] Evaluation Manifesto.
[5] SEU Evaluation Manifesto.
[6] SEU Framework.
[7] As indicated in the Evaluation Manifesto.
[8] For a detailed overview of this process, refer to the SEU Guideline document, The Six-step Process.

quality assurance. Evidence from internal reports[9] suggest the question of quality was heightened in 2018 and 2019 when the unit saw a decrease in evaluation demand across OCB. These and other annual reports document various successes and challenges the SEU has encountered with the support of commissioning, managing, and making use of external evaluations to foster a culture of evaluation for learning and accountability across the organization. Amid all these efforts and the multitude of internal quality assurances through policies, procedures, and oversight from the SEU SC, before this current study, no formal external meta-evaluation had been commissioned for the SEU.

EVALUATION PORTFOLIO

The evaluation dossier or portfolio this meta-evaluation evaluated spans five years from 2017 to 2021 and includes 31 discrete evaluation cases.[10] It is a geographically expansive portfolio, with cases that cover over 20 different countries across six different continents, with many cases including multiple countries, regions, or global settings in one study. It is predominantly focused on medical operations where 68% (n=21/31) cases investigated aspects or complete interventions of medical humanitarian service provision. The other 32% (n=10/31) evaluated non-medical operational projects, policies, and products such as emergency non-food item distributions, novel change management programs, or early warning systems. It represents a diverse cross-section of medical humanitarian operations with over 15 different thematic Transversal or Model of Care types serving as the object of primary evaluations. These included but were not limited to: sexual and reproductive health programming, hurricane responses, victims of torture rehabilitation, refugee assistance, vaccine campaigns against infectious diseases, interventions for non-communicable and chronic diseases, and epidemic response like and COVID pandemic responses. Evaluations that assessed a specific humanitarian or medical humanitarian intervention comprised 71% (n=22/31), emergency evaluations comprised 10% (n=3/31), pilot project/innovation evaluations comprised 6% (n=2/31), and 23% (n=7/31) of total evaluations were coded as organizational-related, that is not directly evaluating a specific medical humanitarian intervention, but components of interventions or systems for the operations department.

Reflecting OCB's policy preference to have evaluations support direct implementation, 71% (n=22/31) of total evaluations were mid-term and only 29% (n=9/31) were end-of-term. The average number of evaluators per evaluation study was two evaluators,[11] with 13 cases having one evaluator only and 3 cases having teams of more than 2 evaluators. Best estimates of average evaluation duration are roughly 6 months.[12] The final evaluation reports had an average of 62 pages, with an average of five annexes, and an average of 15 evaluation questions and sub-questions.[13] The majority of evaluation cases (77%, n=24/31) had additional final report formats such as accompanying "posters" or 1-pagers, and or short-versions. One commissioned evaluation (not included in the total 31 count) was canceled due to limitations related to COVID, and another evaluation case did not publish reports due to a sub-standard final report draft.

Across the portfolio, 97% (n=30/31) of evaluations used at least one OECD-DAC criterion explicitly or implicitly, except for an evaluation that investigated an OCB budget overspend. The average evaluation

[9] SEU quarterly reports for 2018 and 2019

[10] The dossier technically includes one additional evaluation that was canceled before any significant evaluation activities could take place and is not included in the denominator of the high-level description of evaluation attributes.

[11] 1.77 rounded up.

[12] Evaluation duration estimates were based on the date of contract signing and final report publication. Where contracts didn't include dates, the start date on the ToR was used.

[13] The number of evaluation questions are counts of prescribed questions in each ToR.

used a combination of 3 OECD-DAC criteria. The overwhelming majority of evaluations the SEU manages are goal-oriented, with 87% (n=27/31) investigating effectiveness. Other regularly occurring OECD-DAC criteria were Relevance (68%, n=21/31), Efficiency (58%, n=18/31), Impact (52%, n=16/31) with Sustainability (16%, n=5/31) and Coherence (6%, n=2/31) investigated less frequently. Among all evaluations, 94% (n=29/31) used additional evaluation criteria, with the most common criterion being "Appropriateness" occurring (51%, n=16/31) of the time.

The average evaluation contract for external evaluators was €25,225.13, with a total consultant cost of €756,745.00 across 5 years and 30 evaluations.[14] At a high-level, this portfolio covers a considerable amount of human and financial resources for evaluation, and even more when considering the indirect resources of the operations these evaluations investigated.

judgments about the quality of the evaluation portfolio were used to make judgments about the quality of the evaluation system in which the evaluation cases were commissioned, managed, and used.

META-EVALUATION SCOPE

This meta-evaluation was an external retrospective summative meta-evaluation. The object of meta-evaluation was primarily a portfolio of completed evaluation cases and secondarily the evaluation system they were commissioned, managed, and used within. The funding type of this meta-evaluation was a competitive RFP administered and accompanied by the SEU, its steering committee, and a consultation group led by the study co-commissioners. Meta-evaluation activities took place from May to October 2022. Though primary evaluations were conducted across many regions and continents, all meta-evaluation activities were conducted remotely by a distributed team of meta-evaluators based in the US managed by the SEU in Stockholm. The primary intended users of the meta-evaluation are the SEU and the SEU Steering Committee, with secondary intended users being OCB Board, General Direction, and Operations Department.

PURPOSE OF THE META-EVALUATION

The purposes of this meta-evaluation study were to:

1. **Assess Quality and Value:** this meta-evaluation focused on the individual and collective quality of finalized evaluations to OCB managed by the SEU between 2017 and 2021 with findings intended to guide the commission, management, and use of future evaluations.
2. **Establish Coherent Understanding of Valuable Evaluations at MSF:** the intended conceptual and process use of this meta-evaluation is a coherent understanding of what constitutes quality evaluation at OCB through the synthesis of meta-evaluative needs, values, criteria, and standards of evaluation users and existing evaluation policies. This process resulted in an initial synthesized meta-evaluation criteria framework to judge the quality and value of the past 5 years of evaluations and future evaluation processes and products.
3. **Determine Organizational Factors of Evaluation Quality:** where the primary purpose of this meta-evaluation was to describe the quality and value of past meta-evaluations, an important auxiliary purpose is to start the process of explaining those judgments. Or in other words, to

[14] Budget data for one evaluation could not be located.

explore factors that influence the quality and value of evaluations at OCB to strengthen or make changes to existing evaluation processes and products to increase quality and utility within the organizational context.

THE META-EVALUATION QUESTIONS

1. **Define Quality:** What are the most important, relevant, and useful meta-evaluation criteria for MSF in determining the quality and value of past and future evaluations?
   - 1.1. What constitutes credible and actionable evidence for intended users at MSF?
   - 1.2. What values about systematic inquiry are important to intended users at MSF?

2. **Assess Quality:** To what extent do the products and performances of OCB finalized evaluations from 2017-2021 meet generally accepted evaluation quality criteria and standards?
   - 2.1. To what extent do individual and combined evaluations satisfy the Joint Committee Program Evaluation Standards (2011)?
   - 2.2. To what extent do individual and combined evaluations satisfy the ALNAP Quality Proforma (2005)?
   - 2.3. To what extent do the SEU evaluation policies, systems, and aggregate SEU managed evaluations satisfy the UNEG Norms and Standards for Evaluation (2017)?
   - 2.4. To what extent do individual and combined evaluations satisfy MSF and OCB-specific operational and evaluation effectiveness principles?

3. **Estimate Value:** With these meta-evaluative conclusions, to what extent do these completed evaluations provide value to OCB?

4. **Explain Quality:** What factors within MSF organizational sphere of influence mediate the quality of evaluation processes and products and how can MSF use this information to ensure high quality and value evaluations?

THE META-EVALUATION FRAMEWORK

A meta-evaluation framework is a collection of evaluation quality dimensions that is used for comparison to make judgments about the quality of evaluation products and processes. A quality framework conveys how practice *ought* to look like in an ideal sense. In collaboration with the SEU and meta-evaluation consultation group, the meta-evaluation team constructed a custom quality framework of performance criteria, sub-criteria, and indicators of quality through the combination of multiple existing complimentary quality frameworks. This framework is the result of desk review, interviews, focus groups discussions, and consultation with intended users that occurred during the inception phase of the meta-evaluation. It is also technically an answer to the first meta-evaluation question (MEQ1) about defining evaluation quality at OCB, at least for this present study.[15]

[15] Definitional work about evaluation quality is an ever-evolving process in the evaluation profession. There is no reason to believe it shouldn't continually evolve for an institution like the SEU, either. Just because this study found consensus for measurement purposes, doesn't mean conversations about quality have been solved and are over. In fact, the meta-evaluation team hopes this study helps the SEU further clarify what is most important to them from among these quality standards that they would like to emphasize moving forward.

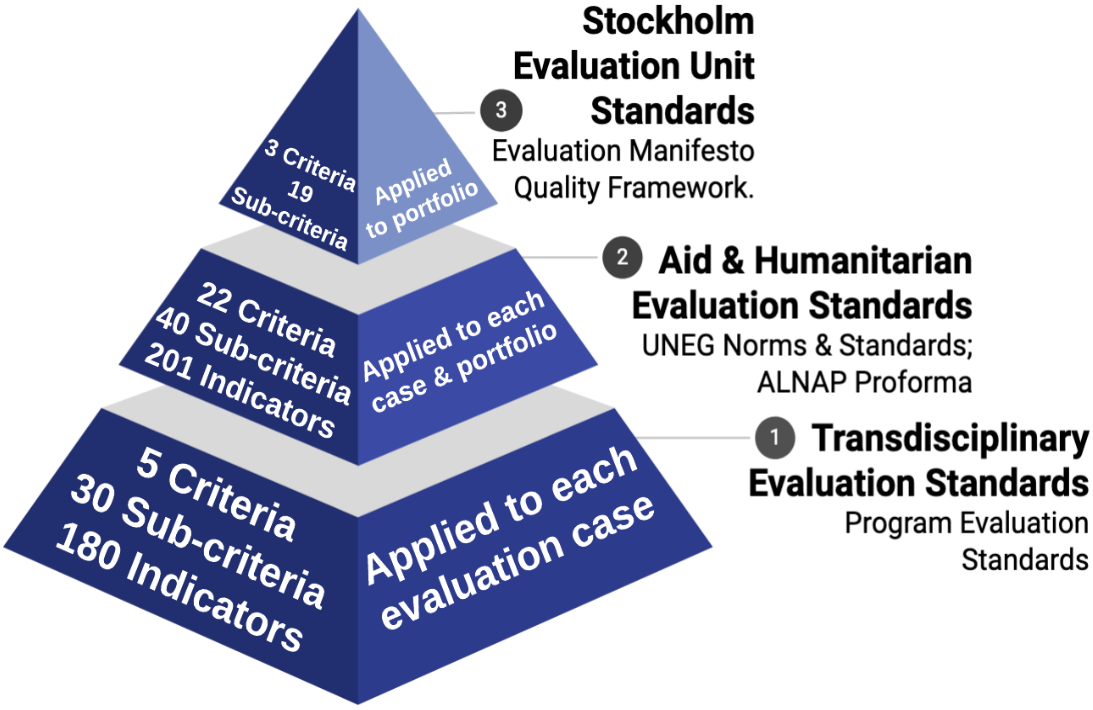


Figure 1: METAE meta-evaluation framework pyramid

The existing quality frameworks that combine into this specific meta-evaluation's framework are 1) *The Program Evaluation Standards* (PrgES); 2) *The United Nations Evaluation Group Norms and Standards of Evaluation* (UNEG); 3) *The Active Learning Network for Accountability and Performance Proforma (*ALNAP); and 4) *The SEU Evaluation Manifesto Quality Framework (EMQ)*.[16] The use of complementary frameworks was based on assumptions that more frameworks: 1) capture a broader notion of evaluation quality; 2) increase the chances intended users will see their specific definition of evaluation quality in the study and be more inclined to trust and use the study; 3) and strengthen the accuracy of and support for overall judgments.

Combined, these frameworks constitute the best available standards for transdisciplinary, international aid, humanitarian action, and SEU-specific evaluation practice.[17] Across the four frameworks there are a total of 32 criteria or dimensions of quality, 89 sub-criteria, 223 specific quality indicators for each case, and 6,913 indicators for the whole portfolio. The following table uses the first criteria, sub-criteria, and 6 indicators of the PrgES framework to demonstrate the relationship between these dimensions of quality.

[16] Due to the absence of existing sub-sectoral quality standards specifically for medical humanitarian evaluation, this domain of quality standards is drawn from the provisional quality framework suggested in the SEU guideline document, The Evaluation Manifesto.

[17] A desk review, expert consultation, and key informant interviews with the SEU were unsuccessful in identifying any quality frameworks for medical evaluation practice, let alone medical humanitarian evaluation practice. Quality of care frameworks were the closest equivalent, but these pertained to medical practice and not medical nor medical humanitarian evaluation practice. For more on this limitation, see the detailed methods note in Annex II.

Table 1. Example criterion, sub-criterion, and indicators

CRITERION	THE UTILITY STANDARDS ARE INTENDED TO ENSURE THAT AN EVALUATION IS ALIGNED WITH STAKEHOLDERS' NEEDS SUCH THAT PROCESS USES, FINDINGS USES, AND OTHER APPROPRIATE INFLUENCES ARE POSSIBLE.					
SUB-CRITERION	U1 Evaluator Credibility. [Evaluations should be conducted by qualified people who establish and maintain credibility in the evaluation context.]					
INDICATOR	Engage evaluators who possess the needed knowledge, skills, experience, and professional credentials	Engage evaluators whose evaluation qualifications, communication skills, and methodological approach are a good fit to the stakeholders' situation and needs	Engage evaluators who are appropriately sensitive and responsive to issues of gender, socio economic status, race, language, and culture	Engage evaluators who build good working relationships, and listen, observe, clarify, and attend appropriately to stakeholders' criticisms and suggestions	Engage evaluators who have a record of keeping evaluations moving forward while effectively addressing evaluation users' information needs	Give stakeholders information on the evaluation plan's technical quality and practicality, e.g., as assessed by an independent evaluation expert

The *Program Evaluation Standards* and the *ALNAP Proforma* frameworks were used to assess all 31 evaluation cases. This translated to 223 unique[18] indicators of quality that were rated "met" or "not met" for each of the 31 evaluation cases for a total of 6,913 qualitative indicators that received their own judgment of quality for the entire evaluation portfolio. The ratios of whether these smallest units of analysis were "met" or "not met" serve as the basis for the quality scores, ratings, and ranks for evaluation cases across the portfolio and for sub-criteria and criteria within cases and across the portfolio.

The *UNEG Norms and Standards* and the *SEU Evaluation Manifesto* framework (EMQF) were not applied to each individual evaluation case, but to the 31 evaluation cases aggregated as a portfolio and to OCB evaluation system. This occurred after initial scores, ratings, and ranks were calculated for the PrgES and ALNAP frameworks.[19] The 168 individual indicators in the UNEG framework were not given their own rating, but the sub-criteria and criteria in both the UNEG and SEU frameworks were given ratings.

Across all frameworks, the sub-criteria and criteria used the same rubric for standards or degrees of quality, which included the following levels: poor, fair, good, very good, and excellent.[20] Summary ratings for all four frameworks are reported in the findings section and detailed scores, ratings, and ranks (where applicable), are included in Annexes 6-9.

[18] Each indicator is uniquely worded, though many indicators, especially between different frameworks, may relate to similar sub-criteria or criteria, such as utility or stakeholder engagement.

[19] A detailed description of the numerical weight and sum methodology used to derive scores, ratings, and ranks for each evaluation case and evaluation criteria is included in the detailed methods note in Annex II.

[20] These standards, or degrees of goodness, are explained in more depth in the detailed methods note.

METHODOLOGY

The following summary[21] of methods and procedures uses the SEU Six-step Process as an organizing structure.

SCOPING

The evaluation client established the meta-evaluation design (retrospective portfolio) and provisional dimensions of quality in the terms of reference—a pared down version of the EMQF, what would become one component of the final meta-evaluation framework. The meta-evaluation team conducted a desk review and offered an initial proposal.

PREPARATORY

This brief stage entailed a series of recruitment interviews, contract negotiations, and an onboarding session. The meta-evaluation team reviewed and signed the Ethical Guidelines document as part of the contracting activity. The meta-evaluation team started taking process notes for participant observation.

INCEPTION

Criteria and standards for evaluation quality were established through an in-depth inception stage that included feedback and discussion with the meta-evaluation manager, interviews with evaluation managers, focus group discussions with various primary and secondary intended users, and dialogue with consultation group members. These recordings were transcribed and analyzed. The framework and meta-evaluation plan were finalized in a detailed inception report, which was approved by the commissioners.

DATA COLLECTION AND ANALYSIS

The PrgES and ALNAP checklists were developed into an online dashboard for data extraction. Two pilot rounds of independent reviewing and comparison were conducted for calibration until inter-rater reliability met acceptable standards. Documents were reviewed and data extracted. Any missing data points were converted into an online survey questionnaire and sent to evaluators and evaluation managers. Projects contacts, commissioners, evaluators, and managers also received questions about experience, satisfaction, and insights into factors of quality. Survey data was added to dashboard and scores, ratings, and ranks calculated for each evaluation case and portfolio for PrgES and ALNAP frameworks. Open-ended survey data was analyzed. Additional metrics were developed from satisfaction scores about evaluations, use and dissemination, and degree of use and influence. Reliability tests were conducted and deemed sufficient. Separate workshops with the SEU and the consultation group were facilitated to discuss processes and set expectations for reviewing the report.

REPORT WRITING

A final report draft was prepared and shared with the evaluation manager who made comments, which the team addressed. The draft report and annexes were then shared with the consultation group and the rest of the SEU who provided feedback. The meta-evaluation team addressed all feedback, made some adjustments to the report, and delivered the final final-report to the evaluation manager.

[21] For more details, see Annex II.

DISSEMINATION AND USE

Evaluation use was a regular discussion point for bi-weekly meetings between the evaluation team and evaluation manager. A workbook on intended users and intended uses was shared with the manager during the data collection and analysis phase to more fully plan for intended use by intended users. A sensemaking session was held with the SEU where findings, conclusions, and recommendations were discussed along with activities for the Use and Dissemination plan. The meta-evaluation team participated in an online webinar for MSF presenting the findings.

ETHICS

While meta-evaluation serves as both a professional and ethical imperative for evaluators, the meta-evaluation team recognized that decisions around its design and management had the potential to cause harm. To mitigate and prevent any harm to any individual during this meta-evaluation, our team engaged with and committed to the ethical values across multiple evaluation quality frameworks including Propriety, Accountability, Integrity, Respect, Common Good, and Equity.[22] In addition, each member of the meta-evaluation team reviewed, signed, and committed to the SEU ethical guidelines before the beginning of the evaluation. Ethical practice included disclosing meta-evaluation purpose and intended use to participants, obtaining informed consent, identifying and mitigating possible harm, and responsible data management. Additionally, it should be noted that the meta-evaluation requester, manager, and central individual of the evaluation system was the same person—the head of the SEU. Addressing potential conflicts of interest related to this unavoidable feature of the meta-evaluation were broached by the head of unit, adequately addressed with the meta-evaluation team, and continually revisited at various points during, inception, data collection, and reporting. This mostly affected issues of when and what to disclose about prior evaluation experiences to avoid undue influence over meta-evaluation findings. The meta-evaluation team attests that these efforts by the head of unit adequately meet the relevant Propriety standards.

LIMITATIONS

This meta-evaluation had a few limitations that were known and unknown at the start of the work. Known issues were tight timeframes for reaching framework consensus; large differences in time-zones between evaluation team members and clients (GMT -10 and GMT +1 at the extremes); lack of French speaker on evaluation team; positionality blind spot in homogenous lived and privileged experiences of evaluators; inter-rater reliability issues without evaluation case dual-rating; and assumed philosophical reservations among some users to the proposed checklist and numerical weight and sum methodologies as potentially too technocratic. Of all of these, time zone differences and timeframes had the most effect in the first four evaluation stages. Timeframe issues were more operational, or more apparent, in data collection and analysis.

The scoped and final meta-evaluation framework was known to be ambitious and even so, level of effort forecasts for data extraction, analysis, and reporting were severely underestimated. This translated to compressed and fast-tracked interpretation and reporting procedures resulting in foregone sophisticated data visualizations in the final report. Instrumentation challenges occurred when pivoting from email interviews to online surveys, which translated into delays and unsatisfied

[22] PrgES; AEA Guiding Principles; UNEG Ethical Guidelines for Evaluation

survey respondents. Self-selection, courtesy, self-serving, recall, and social acceptability biases are all possibilities with the self-report data from our purposefully sampled online survey about past evaluation performances. A survey response rate of 42% (n= 57/133) was decent for an external survey and likely low for an internal survey. Managers comprised the largest respondent group with 42% of total responses (n=24/57), then evaluators with 32% (n= 18/57), followed by commissioners with 16% (n= 9/57) and project contacts with 11% (n= 6/57). Relative to inputs, processes, and outputs, outcome-level data was lacking due to limited evaluation follow-up and use documentation.

Finally, although two subject-matter experts (SME) for medical evaluation and humanitarian evaluation were successfully recruited, engagement with these SMEs was limited and of low influence in the design and execution of the meta-evaluation with almost all input in the reporting stage and ultimately no response and integration of their feedback toward the end of the process when provided with report drafts and technical questions. These and other limitations are addressed in detail in the Annex II. The meta-evaluation team believes these limitations were sufficiently addressed during the conduct and or accounted for in this final report and do not pose undue threats to the validity of meta-evaluation conclusions.

FINDINGS

The ambitious scope of this meta-evaluation translated into a large body of findings. Summary and synthesis of findings was necessary for sensemaking and reporting. The findings section in this main report contains key findings and finding summaries from analysis procedures that correspond to each of the meta-evaluation questions and sub-questions. Meta-evaluation questions about defining quality (MEQ1) and assessing quality (MEQ2) have in-depth corresponding assessment reports and sub-criteria dashboard snapshots in Annexes VI to IX.[23] These annexes and supplemental dashboard file transparently report definitions for all criteria, sub-criteria, and indicators in detail; report all judgments at all levels of these quality dimensions; describe all relevant evidence sources for judgment backings; present detailed descriptions of actual performance and states of evaluation cases and the evaluation system for judgment warrants. Findings about the value or worth of these evaluations to OCB (MEQ3) are presented in this section, along with some corresponding analysis about a key component of worth, use and influence. Finally, findings related to explaining quality assessments (MEQ4) or answering why quality was inadequate, adequate, or exemplary, are offered at the end of this section. See the detailed method note annex for descriptions of the source materials.

[23] A link to the supplemental and detailed dashboard for PrgES and ALNAP case study assessments can be found in Annex V.

MEQ1 (DEFINING QUALITY): WHAT ARE THE MOST IMPORTANT, RELEVANT, AND USEFUL META-EVALUATION CRITERIA FOR MSF IN DETERMINING THE QUALITY AND VALUE OF PAST AND FUTURE EVALUATIONS?

The meta-evaluation terms of reference included a disclaimer drawn from the Evaluation Manifesto that states that the SEU does not “manage a stated quality framework to define what is quality and value in evaluations at OCB” and that “ideas on what constitute quality or value for different stakeholders...differ across the organization.” it further stated that “It will be necessary to establish a framework of accepted criteria as part of the evaluation process.” While the last statement was true, the first two statements are partially true. First, interviews confirmed ideas on evaluation quality differ across interviewees, but these differences are complimentary and not in contradiction with an overall vision of evaluation quality. Second, document content analysis revealed the SEU does in fact manage an (un-stated) quality framework, dispersed across multiple guideline documents.

WHAT VALUES ABOUT SYSTEMATIC INQUIRY ARE IMPORTANT TO INTENDED USERS AT MSF?

Values from MSF and OCB Policy Documents

As a movement, MSF values the evaluation function, or disciplined systematic value-based inquiry, as evidenced by the high-level policy document The La Mancha Agreement. The Agreement states that MSF makes a “commitment to evaluation” and “aspires to ensure quality and relevance in operations, is committed to the impact and effectiveness of its work so that good work can be multiplied and abandon ineffective practice.”[24] The Agreement also acknowledges MSF values accountability and transparency “to those we assist, our donors and wider public.” OCB has translated this high-level commitment to accountability and transparency with a commitment in its 2020-2023 Strategic Orientations[25] to a “culture of evaluation” that “give[s] the field teams the opportunity to learn from [their] practices and to constantly improve the quality and pertinence of operational/medical interventions.” OCB has subsequently adopted an evaluation policy that all projects should be evaluated during their lifespan unless there is a well justified reason not to.

Values from SEU Evaluation Guideline and Policy Documents

‘Values’, ‘Use’, and ‘Method’, are the main dimensions of quality criteria as provided in the Manifesto document. In the table below, the sub-headings provide facets of definitions for each criterion.

Through a content analysis of SEU evaluation policy and guideline documents, the following professional values have been categorized by the SEU’s three conceptual domains of quality (value, use, method) from the emerging framework found in the Evaluation Manifesto.

[24] SEU Evaluation Manifesto

[25] Interestingly, the new SEU Annual Reports function as Principles-focused Evaluations of Strategic Orientations at the OCB. This is a form of formalized meta-evaluation that explores the extent to which effectiveness and operational principles within the Strategic Orientation were manifest in evaluations.

Table 2. Values about evaluation quality from SEU evaluation policy

VALUE	USE	METHOD
Transparency Accountability Downward Accountability Credibility Objectivity Independence Necessity Impartiality Credibility Ethics Honesty and Integrity Inclusivity Engagement and Ownership Respect for Dignity and Diversity Avoidance of Harm	Utility Use Effective Evaluation Processes Learning Real Time Learning Follow up on Findings and Recommendations Culture of Evaluation	Rigor Accuracy Completeness Reliability Confidentiality Quality Assurance/Control

All the stated values here are directly duplicated or associated with criteria, quality domains, and norms and standards in the PrgES, ALNAP, or UNEG Frameworks, which makes sense given the SEU explicitly cites these quality frameworks as inputs to their evaluation policies and conceptions of evaluation quality.

WHAT CONSTITUTES CREDIBLE AND ACTIONABLE EVIDENCE FOR INTENDED USERS AT MSF?

Values from Key Informant Interviews

From interviews and focus group discussions, several key values about systematic inquiry emerged, including inclusiveness, learning and self-reflection, transparency, neutrality and objectivity, relevance, and responsiveness. In addition, values around justice, being “doers not talkers,” “values over technical indicators,” as well as putting “people over profits” emerged as additional descriptions of what was important to MSF.

Key values that pertain to the credibility and utility of evidence for users at OCB include independence, transparency, rigor, and evidence that is grounded in context and insights from various stakeholders, including those less often heard from. Many of these personal and professional values align with generally accepted evaluation quality frameworks and SEU policy documents.

Though the KII protocol did not include an explicit question about the distinctiveness of evidence for medical/health service delivery, “medical ethics” and “Do No Harm” were the top expressions of values pertaining to the evidence of medical health services delivery.

META-EVALUATION FRAMEWORK

Synthesis of preliminary desk reviews, key informant interviews, evaluation policy document analysis, and consultations with the evaluation manager and primary intended users resulted the following hybrid quality framework. This framework—or collection of elements that describe and measure quality evaluation practice—is the most comprehensive answer to this initial meta-evaluation question. It represents how the definitions of evaluation "quality" at OCB are used to measure and make claims about actual evaluation performance managed by the SEU.

Table 3. Detailed composition of METAE meta-evaluation framework

FRAMEWORK	CRITERIA	SUB-CRITERIA	INDICATORS	UNIT OF ANALYSIS	EVIDENCE SOURCE
PrgES	5	30	180	Evaluation Case	Documents; survey; interviews; participant observation
ALNAP	5	15	43	Evaluation Case	Documents
UNEG	19	24	168[26]	Evaluation Portfolio	PrgES and ALNAP Ratings; survey; interviews; participant observation
EMQF	3	12	NA	Evaluation Portfolio	PrgES and ALNAP Ratings; survey; interviews; participant observation
METAE	32	81	391	Case/Portfolio	Documents; PrgES and ALNAP Ratings; survey; interviews; participant observation

MEQ2 (ASSESSING QUALITY): TO WHAT EXTENT DO THE PRODUCTS AND PERFORMANCES OF OCB FINALIZED EVALUATIONS FROM 2017-2021 MEET GENERALLY ACCEPTED EVALUATION QUALITY CRITERIA AND STANDARDS?

Using the meta-evaluation framework, performance scores, ratings, and ranks were determined for each evaluation case and performance criteria across cases for PrgES and ALNAP frameworks. Using these judgments and additional evidence, ratings were then awarded to the entire portfolio for norms, standards, and other quality domains for the UNEG and EMQF frameworks. Summary scores, ratings, and ranks for each framework are reported in this section along with key findings and key recommendations for improvement. All judgments and conclusions were made by the external meta-evaluation team.

OVERALL PROGRAM EVALUATION STANDARDS RATING: GOOD (60%)

The portfolio of evaluations managed by the SEU from 2017-2021 received a GOOD rating with a score of 60% when compared with the highest bar of transdisciplinary evaluation quality. To an uninformed

[26] Judgments were made at the sub-criterion level for this framework. Indicators were scanned and considered, but judgments on whether specific indicators were met or not met were not made for this framework like the PrgES and ALNAP frameworks.

reader these findings may seem underwhelming. Despite the room for improvement, this is a laudable result. A total of 5,580 judgments across 31 evaluation cases about evaluation processes and products informed by a constellation of evidence[27] combine into this overall portfolio score and rating. This finding reveals OCB has a healthy and well-functioning evaluation institutional framework and emerging evaluation culture and suggests high value of evaluations for OCB.

Table 4. PrgES Criteria Ranks, Scores, and Ratings

Utility		Feasibility		Accountability		Propriety		Accuracy		Total Quality	
Score	Rating	Score	Rating	Score	Rating	Score	Rating	Score	Rating	Score	Rating
68%	Very Good	66%	Good	63%	Good	53%	Good	51%	Good	60%	Good

1. UTILITY: VERY GOOD (68%)

Criterion Definition: The *UTILITY* criterion examines the extent to which evaluations are "aligned with stakeholders' needs related to contribution towards use and influences are possible." Emphasis should be placed on the word "possible" use and influence (results of use) as this criterion does not directly measure actual evaluation use and influence. This criterion relates to MSFs overall pragmatic culture of being "doers, not talkers."

SEU Description: The highest scored and subsequently rated and ranked criterion across all evaluation cases for the SEU portfolio is *UTILITY.* The SEU has strong evidence[28] of "Very Good" utility, meaning the potential for evaluation use is high for OCB. This is not by accident, but a result of intentional and concerted efforts by the head of the SEU, evaluation managers, qualified external evaluation consultants, invested consultation groups, supportive evaluation commissioners, and helpful project contacts. It is evidence of a long-term strategy to foster evaluation culture at Operational Centre Brussels. This is a promising finding for OCB which places high value on the practicality and usefulness of information for decision-making. The SEU has *Meaningful Processes and Products* (97%), which means evaluation processes and products regularly and consistently meet the needs of evaluation participants and encourage use. The SEU produces *Relevant Information* (88%) that meets the planned for and emergent information needs of intended users. The SEU has *Timely and Appropriate Communication and Reporting* (84%) that meets dissemination needs of intended users and most right-to-know audiences and is a result of specific evaluation policies, procedures, and good management. Within this high-performing criterion, there are still some areas for improvement, which are described below with attendant recommendations for improvement.

Recommendations: the SEU can do more to increase *Evaluator Credibility* (53%), specifically developing policies and procedures that ensure the appropriate evaluation participants are provided information on "the evaluation plan's technical quality and practicality" potentially assessed by an independent evaluation expert (0%). Further, the SEU can improve descriptions of *Explicit Values* (54%) in their evaluation reports that "clarify and specify the individual and cultural values underpinning the evaluation purposes, processes, and judgments" especially given the cardinal importance of *Values* in

[27] General evidence sources for indicator judgments across PrgES and ALNAP frameworks are evaluation artifacts, online survey data, key informant interview and focus group discussion transcripts, and participant observation notes.

[28] Specific evidence for this criterion includes, but is not limited to, 48 unique qualitative indicators across 31 cases for a total of 1,488 judgments for the portfolio.

the EMQF. Related to *Attention to Stakeholders* the SEU systematically underperforms at "Search[ing] out and invite[ing] input from groups or communities whose perspectives are typically excluded, especially stakeholders who might be hindered by the evaluation" (10%). This mostly absent practice contradicts the effectiveness principle from the EMQF to *Engage the Voices of Those Less Present.* Specific procedures to fulfill this policy need to be reviewed, updated, or enacted to close this gap. Finally, related to the SEUs *Concern for Consequences and Influence,* the SEU needs to recommit to "Follow up [with] evaluation reports to determine if and how stakeholders applied the findings" (10%) and measure the extent to which these applications lead to specific results for operations and potentially for intended beneficiaries of interventions being evaluated.

2. FEASIBILITY: GOOD (66%)

Criterion Definition: The *FEASIBILITY* criterion examines the extent to which evaluations are "viable, realistic, contextually sensitive, responsive, prudent, diplomatic, politically viable, efficient, and cost effective."

SEU Description: Ranked at a close second, the SEU has strong evidence[29] of "Good" *FEASIBILITY* relative to the other criteria. This finding is indicative of high-quality evaluation management in accordance with robust policies and procedures. The SEU equips evaluation managers and consultants with a clear vision of their view of the purpose of evaluation and translates this vision into actionable procedures. The head of unit accommodates for reasonable variation in application of these procedures according to individual interpretation, emphasis, and unique skill sets across the team of evaluation managers and coordinator. The SEU has "Very Good" *Project Management* strategies (78%), *Practical Procedures* (77%), and *Resource Use* (67%). Of importance, evaluations managed by the SEU were *Contextual Viability* (79%), and consistently "recognize, monitor, and balance the cultural and political interests and needs of individuals and groups." The SEU accomplished this through detailed background and contextual analyses in evaluation reports, mandated consultation group formation and accompaniment, and appropriate mechanisms for participants to stay informed about the evaluation findings and progress.

Recommendations: within this criterion, the SEU can make the most advancements in "Assess[ing] and confirm[ing] the program's evaluability before deciding to proceed with the evaluation" (23%). This finding reveals a critical quality gap related to the EMQF effectiveness principle of *Consider the Evaluability of the Project.* The main source of evidence for this gap was select evaluation reports that scoped specific evaluation criteria and questions, and due to data availability and other related (reasonably foreseeable) limitations, were unable to answer key evaluation questions or answer them sufficiently, mostly related to outcomes and impact. This consequence of sub-standard evaluability assessment procedures was not widespread, but reoccurring. Limited scoping questions documents (n= 8/31) and pre-natal documentation (not reviewed) provide evidence that evaluability is considered. Other artefacts that had limited use in the 2017-2021 period were stakeholder mapping and situational analyses documents. The SEU should consider revising and consolidating elements of all active and inactive artefacts into a revised scoping workbook and checklist that includes new dimensions related to existing work on evaluability assessments.[30] Finally, and of significant consequence to the SEU and OCB is that not one evaluation systematically "Document[ed] the

[29] Specific evidence includes 24 unique qualitative indicators across 31 cases for a total of 744 data points for the portfolio.

[30] See Annex XII for a list of recommended resources.

evaluation's benefits, including contributions to program improvement, future funding, better informed stakeholders, and dissemination of effective services." Reasons for this omission in evaluation follow-up need to be discussed and addressed internally as this is not only a generally accepted standard, but also an effectiveness principle from the EMQF, *Follow Up on Findings and Recommendations.*

3. EVALUATION ACCOUNTABILITY: GOOD (63%)

Criterion Definition: The *EVALUATION ACCOUNTABILITY* criterion examines the extent to which an evaluation is systematically, thoroughly, and transparently documented and then assessed, both internally and externally for its utility, feasibility, propriety, and accuracy. This criterion refers to the regular practice of meta-evaluation—evaluating evaluations—formal and informal, internal and external. This aligns with OCB commitment to accountability and learning applied to evaluation, not just operations.

SEU Description: Ranked third, moderately strong evidence[31] of a "Good" rating for this criterion is an indication that the SEU practices what it preaches to medical humanitarians, namely that its own professional practice— medical humanitarian evaluation—is evaluated. Meta-evaluation within the SEU takes the shape of regular quality assessment checkpoints throughout the six-step process of evaluation. Statements of evaluation quality across guideline documents, prospective evaluation consultant score cards, proposal reviews, interviews, inception report reviews by consultation group, bi-weekly meetings between evaluators and managers, sensemaking session and workshops, and final report reviews all provide in-built warrants for and moments of quality assessment of evaluation processes and products. The SEU received a perfect score and "Excellent" rating for *External Meta-evaluation* (100%) based on the commission and known procedures of this meta-evaluation study. External portfolio meta-evaluation is a rare practice, and this rating is a recognition of the SEU's commitment to evaluation accountability. This was awarded despite the fact no individual evaluation case had an external meta-evaluation conducted. Given the average evaluation budget, and considerable internal evaluation capacity to assess quality, the estimated return on investment of formal external meta-evaluations for each case is questionable. Aspects of *Evaluation Documentation* (74%) scored near "Excellent" across the board except for "Evidence of the evaluation's consequences, including stakeholders' uses of findings" (0%). Recommendation follow-up documents existed for three evaluation cases but were deemed insufficiently used.

Recommendations: though internal quality assessment processes are present, the biggest area for improvement is *Internal Meta-evaluation* (50%). First, the SEU needs to match its strong perspective about the nature of evaluation with a strong perspective about evaluation quality—and own it. Building blocks for an institutional meta-evaluation framework exist across SEU policy documents and culminate in the EMQF. However, the SEU maintains it "[does] not manage a stated quality framework" as a pre-amble to the description of what is arguably an emerging and limited quality framework in the Evaluation Manifesto. The EMQF is a solid start for the SEU and the meta-evaluators recommend updating the framework in the following ways: 1) Consider reframing all dimensions of evaluation quality as effectiveness principles, strategic and operational; 2) consolidate additional descriptions of evaluation quality in other guideline documents to the Evaluation Manifesto and accompanying EMQF; 3) add an additional dimension of *Transformation* that houses principles related to culturally

[31] 18 unique qualitative indicators across 31 cases for a total of 558 data points for the portfolio.

responsive and equitable evaluation that attend to issues of social justice and human rights;[32] 4) consider conducting an internal study or joint exercise that determines if medical humanitarian evaluation requires specialized and additional notions of evaluation quality. Answer the question of "what are the exceptional limitations or opportunities that dictate an adapted approach to evaluation practice at the SEU versus for international development, or some other non-humanitarian social amelioration endeavor?" Second, the SEU should update procedures to match updated evaluation quality policies. Identify which critical milestones could receive additional support. Areas for consideration are *Evaluability Assessment* work in the Scoping stage and *Evaluation Follow-Up* in the Dissemination and Use stage. The meta-evaluation team recommends the use of rubrics and checklists to improve transparency, accuracy, and reliability of this internal meta-evaluative function.

4. PROPRIETY: GOOD (53%)

Criterion Definition: The *PROPRIETY* criterion examines the extent to which evaluation are conducted properly, fairly, legally, ethically, and justly with respect to (1) evaluators and stakeholders' ethical rights, responsibilities, and duties; (2) systems of relevant laws, regulations, and rules; and (3) roles and duties of professional evaluators. This criterion speaks to the SEU's special interest and value in ethical evaluation practice. It is also the main, but not only domain, that pertains to transformative, equitable, and culturally responsive evaluation practice.

SEU Description: Fourth in overall performance rank, *PROPRIETY* performance in the SEU portfolio has strong evidence[33] of a "Good" rating and suggests the SEU's managed external evaluations are ethical, with room for improvement. While the SEU gets many things right in this criterion, this score and ranking are indicative of a minor quality gap between evaluation policy and practice. Items in the ToR of this meta-evaluation, discussions in the planning and inception stages, the Ethics sub-domain in the EMQF, and the Ethical Guidelines document (which the SEU now requires all consultants to read and sign before starting activities) suggest a high value on propriety in evaluation practice and management. However, the *Propriety* criterion was ranked lowest in importance of all other PrgES criteria by members of the meta-evaluation consultation group during a virtual workshop with the evaluation team in early October 2022. Further, some key standards were not being met. In terms of what the SEU does right, the SEU has "Excellent" *Formal Agreements* (100%), that are "negotiated to make obligations explicit and take into account the needs, expectations, and cultural contexts of clients and other stakeholders." Further, the SEU is "Very Good" at *Transparency and Disclosure* (76%), meaning an overwhelming majority evaluations managed by the SEU "provide complete descriptions of findings, limitations, and conclusions to all stakeholders unless doing so would violate legal or propriety obligations." Finally, SEU managed evaluations are "Very Good" at *Responsive and Inclusive Orientation* (69%), meaning most evaluation cases are responsive to the needs of evaluation participants and communities by taking into consideration the interventions context, gathering useful information from evaluation participants, and designing multiple opportunities for participants to be involved. These are met through the useful policy standard of mandating a consultation group for every case and allocating in-depth time for inception stage activities.

Recommendations: the SEU can make significant gains with *Human Rights and Respect* (51%). Specifically, the SEU should *"develop and communicate rules that assure fairness and transparency in*

[32] For resources about principles-focused evaluation and culturally responsive and equitable evaluation see Annex XII.

[33] 42 unique qualitative indicators across 31 cases for a total of 1,302 data points for the portfolio.

deciding how best to allocate available evaluation resources to address the possible competing needs of different evaluation stakeholders" (0%). Also, in efforts to *Engage the Voices of Those Less Present* the SEU should, *"before releasing the evaluation's findings, inform each intended recipient of the evaluation's policies—regarding such matters as right-to-know audiences, human rights, confidentiality, and privacy— and, as appropriate, acquire her or his written agreement to comply with these policies"* (32%). Finally, pre-empting and complimenting the signing of the SEU Ethical Guidelines, the SEU can include an item in the ToR that requires evaluation teams include *"the evaluator's ethical principles and codes of professional conduct…"* (42%) in their proposals, and to continue this disclosure process in inception reports, informed consent procedures, and final reports.

5. ACCURACY: GOOD (51%)

Criterion Definition: the *ACCURACY* criterion examines the extent to which evaluations "[employ] sound theory, designs, methods, and reasoning in order to minimize inconsistencies, distortions, and misconceptions and produce and report truthful evaluation findings and conclusions." This criterion aligns with stated values of SEU and OCB interviewees and evaluation policies around rigor, objectivity, credibility, reliability, and completeness and the EMQF domain of *Method.*

SEU Description: Ranked last, the *ACCURACY* criterion still received a "Good" rating on the low end of that standard with a moderate score of 51% with strong evidence for these results.[34] These findings signal SEU's attention to the *Method* domain from the SEU Manifesto Quality Framework and confirm to OCB that evaluation findings and conclusions are reasonably sound and support an assumption that application of recommendations is warranted. The SEU does a "Very Good" job of *Explicit Program and Context Descriptions* (89%), with most reports providing detailed accounts of evaluation objects. The SEU also does "Very Good" at *Communicating and Reporting* (88%) Evaluations reports and communications from the SEU have adequate scope and guard against misconceptions, biases, distortions, and errors. Finally, SEU evaluations regularly have *Sound Designs and Analysis* (72%), meaning evaluations "employ technically adequate designs and analyses that are appropriate for the evaluation purposes."

Recommendations: as this is the lowest scoring criterion, there are more opportunities for improvement. First, gains can be made in specific standards of *Information Management* (35%). It should be noted that this sub-criterion received low scores mostly due to a lack of evidence that these standards were met in documentation and survey responses, which likely corroborates this finding. The SEU should review procedures related to "document[ing] and maintain[ing] both the original and processed versions of obtained information; retain[ing] the original and analyzed forms of information as long as authorized users need it; store[ing] the evaluative information in ways that prevent direct and indirect alterations, distortions, destruction, or decay." Additionally, *Reliable Information* (41%) can be improved with the following recommendations: require evaluators to discuss in final evaluation reports 1) *"the consistency of scoring, categorization, and coding and between different sets of information, e.g., assessments by different observers*." And encourage evaluators to "report the needed types of reliability—e/g., test-retest, findings from parallel groups, or ratings by multiple observers—and the acceptable levels of reliability" in inception reports. Finally, and of significant importance the SEU should expect and promote improved *Explicit Evaluative Reasoning* (45%) in both primary evaluations and quality assessments for internal meta-evaluation. Given the SEU is moving

[34] 48 unique qualitative indicators across 31 cases for a total of 1,488 data points for the portfolio.

toward more collaborative sensemaking and recommendation generation, final reports should "identify the persons who determined the evaluation's conclusions [and recommendations], e.g., the evaluator using the obtained information plus inputs from a broad range of stakeholders" (0%). Policies and reporting templates can also be updated to encourage 1) "report[ing of] plausible alternative explanations of the findings and explain why rival explanations were rejected" (0%); 2) "[examination] and report[s of] how the evaluation's judgments and conclusions are or are not consistent with the possibly varying value orientations and positions of different stakeholders" (3%); and 3) "Identify, evaluate, and report the relative defensibility of alternative conclusions that might have been reached based on the obtained evidence" (3%). With such a strong emphasis on "valuing" in evaluation policy documents, a significant number of evaluations can be more transparent and explicit in documenting the evaluations chain of reasoning, which includes what standards are being used to make evaluative judgments. This is likely one of the most significant gaps of evaluation policy and procedures: the SEU prescribes and negotiates appropriate criteria for all evaluations, but there is no evidence of any discussion of standards or degrees of those criteria are established, let alone discussions of which degrees are acceptable or satisfactory for specific criteria. A practical example of this, which the meta-evaluators hope to discuss, is determining if the scores and ratings for criteria and frameworks for this meta-evaluation are acceptable or not acceptable to the SEU and OCB.

OVERALL ALNAP PROFORMA RATING: VERY GOOD (80%)

When compared with the best available and generally accepted sector-specific quality framework for humanitarian evaluation, strong evidence reveals the portfolio of past SEU managed evaluations is "Very Good" with a score of (80%). This framework makes direct claims about the quality of evaluation reports, which provide indirect claims about the quality of evaluation processes. With that, these results indicate the SEU manages the delivery of strong evaluation reports that attend to specific considerations of humanitarian evaluation practice. While this meta-evaluation made comparisons between evaluation cases and sub-portfolios by year, these sector-specific ratings can be used to make comparisons with similarly sized centralized evaluation units and evaluation portfolios from other medical humanitarian institutions. While the meta-evaluation team only had a cursory glance at publicly available meta-evaluations on ALNAP, we are confident the SEU and OCB would be in the upper percentile of results for comparable portfolios and units.

Table 5. ALNAP Proforma Domain Ratings, Scores, and Ratings

Report Assessment		Terms of Reference		Contextual Analysis		Intervention Assessment		Methods		Total Quality	
Score	Rating	Score	Rating	Score	Rating	Score	Rating	Score	Rating	Score	Rating
88%	Very Good	82%	Very Good	80%	Very Good	78%	Very Good	72%	Very Good	80%	Very Good

1. REPORT ASSESSMENT: VERY GOOD (88%)

Domain Definition: This quality domain assesses specific dimensions of evaluation report comprehensiveness and quality including the *Findings, Conclusions and Recommendations* and *Report Coverage, Legibility, and Accessibility*.

SEU Description: *Report Assessment* had strong evidence[35] the highest scored and ranked quality domain for this framework, and rightfully so. The SEU has templatized their evaluation reports with useful guidance for consultants to consider when presenting evaluation results. The template requires major information elements but allows enough license for practitioners to present results in ways matched for evaluation cases, contexts, and intended uses for intended users. The SEU manages evaluations that produce "Very Good" *Report Coverage, Legibility, and Accessibility* (88%) and "Very Good" *Findings, Conclusions and Recommendations* (91%). Most evaluation cases have multiple reporting formats like final reports, short versions, and 1-pager "posters." Most evaluations are highly readable prioritizing essential information using executive summaries, right-sized findings sections, and judicial use of annexes for deeper interrogation of the appropriateness and credibility of evidence and methods.

Recommendations: this domain presents little room for improvement against this specific rubric of report quality, but from this framework the *Executive Summary* (80%) function presents the most room for improvement. Specifically, a few cases of executive summaries were deemed insufficient, not summarized enough, or omitted key findings and recommendations that should have been included.

2. TERMS OF REFERENCE: VERY GOOD (82%)

Domain Definition: This quality domain addresses the extent to which the terms of references include all relevant and comprehensive information deemed necessary to establish the scope of the evaluation and attract the proper talent to meet evaluation needs. This includes specific *Terms of Reference* elements and *Expectations of Good Evaluation Practice*.

SEU Description: Ranked second with a score of 82%, this domain was rated "Very Good" with strong evidence.[36] There is variation in Terms of Reference through the years, with the most recent documents exhibiting higher comprehensiveness. Some terms of reference are more detailed than others in certain domains. There is a decent degree of balance between description and prescription in these scoping documents and additional information from inception reports reveals terms of references are the starting point for a dialogue and discussion for further scoping considerations with the external evaluation consultants. The SEU does an "Excellent" job of establishing *Expectations of Good Evaluation Practice* (100%). Almost every ToR references some or all of the OECD-DAC, which is promising. Many reference SEU guiding documents. Onboarding meetings during the planning stage are moments where these values and expectations are likely reiterated (they were for this meta-evaluation). The meta-evaluation team and consultants are required to review and sign the Ethical Guidelines since its publication.

Recommendations: two specific areas for improving the Terms of Reference are 1) more fully describing the nature of the evaluator selection process (e.g., competitive bidding, standing offer) (23%), though ToRs improved on this in recent years, the nature should be explicit, 2) and ensuring the SEU has fully articulated the "rationale for the timing of the evaluation" (61%).

[35] 7 unique qualitative indicators across 31 cases for a total of 217 data points for the portfolio.

[36] 7 unique qualitative indicators across 31 cases for a total of 217 data points for the portfolio.

3. CONTEXTUAL ANALYSIS: VERY GOOD (80%)

Domain Definition: This quality domain addresses the extent to which the evaluation reports conduct adequate analysis of the humanitarian context. Two main aspects of this domain are 1) the analysis of the context and of the crisis to which the intervention is responding and 2) Past involvement of the agency and its local partners.

SEU Description: there is sufficient evidence[37] to demonstrate that most evaluations contained sufficient and exemplary context assessments in the reports that provide an adequate amount of information for audiences to situate evaluation findings. Many provide timelines and in-depth detail. In some instances, an assessment of a specific crisis and its history and social setting is not applicable, but when it is, most SEU evaluations meet this bar handedly.

Recommendations: the SEU should work with the steering committee to establish acceptable levels of analyses of the background and context of humanitarian crises for all intended users and right-to-know audiences. What is good enough and how might these contextual analyses conducted presumably in the inception phase inform the balance of the evaluation activity?

4. ASSESSING THE INTERVENTION: VERY GOOD (78%)

Domain Definition: This quality domain addresses the extent to which the evaluation report describes and assesses the humanitarian intervention, which includes 1) *Institutional Considerations*; 2) *Needs Assessment, Objectives, Planning, and Implementation*; 3) *Application of Evaluating Humanitarian Action (OECD-DAC) Criteria*; and 4) *Consideration given to Cross-cutting Issues.* The breadth of this subdomain makes it the most accurate measure of all the ALNAP domains.

SEU Description: strong evidence[38] demonstrates the SEU manages the delivery of reports that do "Very Good" at *Assessing the Intervention,* or object of primary evaluation. These reports attend to important *Institutional Considerations* like the agencies guiding principles and the agencies management and human resources. Reports are "Very Good" at describing needs assessments, objectives, and program cycle processes. The reports also do "Very Good" at the *Application of EHA Criteria.* Many of these results are due to the standardization of evaluation reporting at SEU with detailed inception reports and final reports. Ten out of 31 evaluation cases, or 32% of the portfolio, had at least one criterion that was deemed insufficiently applied. Out of the total 118 intentional attempts to apply OECD-DAC criteria, 17% (n=20/118) were deemed insufficient. This means that evaluation cases had instances of evaluation questions not being able to rate OECD-DAC criteria or had ratings that were deemed not sufficient.[39] These were mostly due to data availability limitations in the lack of outcome-level data for operations. Finally, and of significant importance, various cross-cutting issues were not sufficiently investigated such as *Advocacy* (58%), *The use of and adherence to international standards* (48%), and *Gender Equality* (32%).

Recommendations: More consideration should be given to mitigation strategies for known evaluability issues related to data availability and quality. While finite resources do affect the extent to which these issues are investigated, assuming these are priorities for the SEU, remediation for these

[37] 3 unique qualitative indicators across 31 cases for a total of 93 data points for the portfolio.

[38] 18 unique qualitative indicators across 31 cases for a total of 558 data points for the portfolio.

[39] Evaluation cases that intentionally did not apply some OECD-DAC criteria were not rated negatively because of this feature, only those that attempted to apply some criteria and were unable to due to unmitigated limitations.

performance gaps can be achieved through updated evaluation report templates for both inception and final reports. In addition, many of these issues are likely to be covered with the overall adoption of evaluation policies that are more explicitly culturally responsive and equitable.

5. METHODS ASSESSMENT: VERY GOOD (72%)

Domain Definition: This quality domain investigates the strength of the evaluation design and methodology as described in evaluation documentation and includes aspects such as the 1) *Nature, make up and appropriateness and biases of the evaluation team*; 2) *Clarification process*; 3) *Appropriateness of the overall evaluation methods*; 4) *Consultation with and participation by primary stakeholders*; 5) *The use of and adherence to international standards*; and 6) *Evaluation constraints.*

SEU Description: although the lowest ranked quality domain, this domain still received a rating of "Very Good" with the score of 72% with moderately strong supporting evidence.[40] There is also an easy explanation for the reduced score for this portfolio. The findings for this domain were greatly affected by a low score in *Nature, make up and appropriateness and biases of the evaluation team* (5%) skewing the results.[41] The other seven sub-domains where received "Very Good" ratings and one (*Evaluation Constraints)* received "Excellent." Aside from this specific gap, these results show SEU managed evaluations that had appropriate and defensible evaluation methods.

Recommendations: update Inception report and final report templates to include a section in the body or annex where evaluation teams can address their composition and potential biases. This is most important for the conduct of the evaluation to be included in the inception report, and then for readers to make inferences about credibility of findings in the final report.

UNEG NORMS AND STANDARDS RATING: GOOD

This meta-evaluation produced strong evidence that the SEU evaluation portfolio and evaluation system perform "Good" against one of the most generally accepted quality frameworks for evaluating international aid evaluation cases and systems, the UNEG Norms and Standards. Of all the quality frameworks, this one deals the most with the institutional framework and evaluation culture that backdrop evaluation case performance. Annexes VI and VII contains detailed definitions, descriptions, and recommendations for all Norms and Standards. This section includes summary tables for Norms and Standards and key findings analyses of the top three and bottom three Norms as well as the top three and bottom three Standards. All judgments were made by the external evaluation team. After initial report feedback from the SEU primary intended user group, description, and ratings for UNEG Norm 4, Evaluation Use and Follow Up were updated from "Poor" to "Fair" after a second round of data analysis of use data. Follow-up remained poor, but Use was rated as "Good" and the Norm which combined the two constructs received the rating "Fair."

Overall UNEG Norms Rating: GOOD

When compared against the list of General Norms, or evaluation quality criteria and domains, the SEU receives "Very Good" ratings for important criteria such as *Utility, Credibility, Independence,*

[40] 8 unique qualitative indicators across 31 cases for a total of 248 data points for the portfolio.

[41] If this sub-domain is not included in the domain average, the domain score increases to 85% and this domain is re-ranked as second overall. This demonstrates the sensitivity of scores for ALNAP domains that have at their most 18 indicators and at their fewest, 3 indicators.

Transparency, Professionalism, Evaluation Policy, and *Responsibility for the Evaluation Function.* At a high-level, these confirm OCB has a healthy and robust evaluation function maintained by a team of competent managers, coordinator, and head of unit. The SEU does a "Good" job of attending to *Ethics* and attempting to promote an *Enabling Environment* through a culture of valuing evaluations at OCB. Many of these Norms overlap with other criteria and quality domains across the other three quality frameworks and receive corresponding ratings. These include the lowest rated domains of *Human Rights and Gender Equality* (Fair) and *Evaluation Use and Follow-up* (Poor). The "Fair" rating for *Impartiality* is also useful finding for the SEU and OCB to consider that complements other frameworks.

Table 6. UNEG Norms Definitions and Ratings

NORM	DEFINITION EXCERPT	RATING
Utility	"…there should be a clear intention to use [the evaluation] to inform decisions and actions"	VERY GOOD
Credibility	"Credibility is grounded on independence, impartiality and a rigorous methodology."	VERY GOOD
Independence	"…evaluators [should] be impartial and free from undue pressure throughout the evaluation process."	VERY GOOD
Transparency	"Evaluation products should be publicly accessible."	VERY GOOD
Professionalism	"Evaluations should be conducted with professionalism and integrity."	VERY GOOD
Evaluation Policy	"clear explanation[s] of the purpose, concepts, rules and use of evaluation within the organization"	VERY GOOD
Responsibility for the Evaluation Function	"governing body [is] responsible for…independent, competent and adequately resourced evaluation [unit]"	VERY GOOD
Ethics	"integrity and respect for [culture]…human rights… gender equality; and 'do no harm'"	GOOD
Enabling Environment	"an organizational culture that values evaluation as a basis for accountability, learning and evidence-based decision-making"	GOOD
Impartiality	"The key elements of impartiality are objectivity, professional integrity and absence of bias."	FAIR
Human Rights and Gender Equality	"[integrate] principles of human rights and gender equality… into all stages of an evaluation."	FAIR
Evaluation Use and Follow-up	"…promote evaluation use and follow-up, using an interactive process that involves all stakeholders."	FAIR
OVERALL NORM RATING FOR SEU EVALUATION PORTFOLIO and SYSTEM		GOOD

1. PROFESSIONALISM: VERY GOOD

Norm Definition: "Evaluations should be conducted with professionalism and integrity. Professionalism should contribute towards the credibility of evaluators, evaluation managers and evaluation heads, as well as the evaluation function. Key aspects include access to knowledge; education and training; adherence to ethics and to these norms and standards; utilization of evaluation competencies; and recognition of knowledge, skills, and experience. This should be supported by an enabling environment, institutional structures, and adequate resources."

SEU Description: This meta-evaluation did not explicitly draw on any evaluator competencies framework, nor were the SEU personnel the primary object of meta-evaluation. Nevertheless, there is strong evidence[42] for a "Very Good" rating of *Professionalism* indicating the SEU is staffed by knowledgeable, skilled, and ethical professionals. Our participant observation and experience in being managed by the SEU largely informs our eventual rating of this norm. The SEU is aware of and attempts to embody important evaluator competencies and references the American Evaluation Association evaluator competencies in their policy documents and purports to follow a competency-based selection process. Evaluation policy and practice points are debated, codified, and enacted in the unit through supportive evaluation management and accompaniment. The *UTILITY* sub-criterion about credibility was judged "good" along with the *PROPRIETY* criterion. The *FEASIBILITY* criterion is 2 percentage points shy of the "very good" cut score. Overall portfolio ratings suggest the SEU has been able to attract and contract external evaluators with a high degree of professionalism.

Recommendations: consider small changes to regular SEU meetings to foster ongoing professional development moments such as having a rotating schedule of managers doing a brief show and tell moment at the start of meetings highlighting a specific evaluator competency from the AEA competencies framework or guiding principles, the Program Evaluation Standards, UNEG Norms and Standards, or MSF-specific sensitizing principles. Consider adding individual professional development plans for evaluation managers to annual plans where managers and the head of unit take individual self-assessments of evaluator competencies and make goals and plans to improve in core competencies with regular informal or formal check-ins about professional progress.

2. RESPONSIBILITY FOR THE EVALUATION FUNCTION: VERY GOOD

Definition: "An organization's governing body and/or its executive head are responsible for ensuring the establishment of a duly independent, competent, and adequately resourced evaluation function to serve its governance and management needs. The evaluation budget should be commensurate to the size and function of the organization. The governing body and/or the executive head are responsible for appointing a professionally competent head of evaluation and for fostering an enabling environment that allows the head of evaluation to plan, design, manage and conduct evaluation activities in alignment with the UNEG Norms and Standards for Evaluation. The governing body and/ or the executive head are responsible for ensuring that evaluators, evaluation managers and the head of the evaluation function have the freedom to conduct their work without risking their career development. Management of the human and financial resources allocated to evaluation should lie with the head of evaluation in order to ensure that the evaluation function is staffed by professionals with evaluation competencies in line with the UNEG Competency Framework. Where a decentralized

[42] Participant Observation; PrgES U1, Feasibility, Propriety.

evaluation function exists, the central evaluation function is responsible for establishing a framework that provides guidance, quality assurance, technical assistance, and professionalization support."

SEU Description: Strong evidence[43] suggests that the *Responsibility for the Evaluation Function* at OCB is "Very Good." The SEU is a centralized evaluation unit—for Operational Centre Brussels—that does not conduct evaluations but manages the conduct of external evaluation consultants. The unit is run by a professional, skilled, and ethical head of unit that has been leading efforts to consolidate evaluation policy, increase institutional reputation, and promote a broad culture of evaluation among primary evaluation users and prospective users. Sorting evaluation cases by year reveals annual increases in evaluation quality over the past five years. The head of unit oversees a team of trained and capable managers, equipped with different backgrounds, strengths, and capacities. No evidence exists to suggest that the head of unit or managers are unable to fulfill their responsibilities by limitations caused by their organizational structure or position within OCB. Analysis of survey data suggests that managers have a very detailed and thorough understanding of their responsibility in ensuring evaluation quality, both for specific cases and in general, at all stages, but especially in the scoping, preparation, and inception stages. Survey responses from evaluation commissioners about their specific role in ensuring the quality of evaluations cite the need for more responsibility in scoping and dissemination and use stages. This recognition coupled with performance data on use and follow-up suggest greater accountability measures for commissioners may need to be in place to ensure the potential value of high-quality evaluations are being realized. Limited project contact data about roles and responsibilities suggest greater importance on data collection and analysis in terms of serving as a link between evaluation and operational teams. The evaluation coordinator did not respond to the question about roles and responsibilities in the survey by design, but interview data suggests a clear understanding of the role. The meta-evaluation team is aware of efforts to shift administrative tasks to MSF Sweden human resources to free up SEU managers and administrators to focus on promoting use and dissemination efforts as well as transversal learning.

Recommendation: With a recent loss of an evaluation manager, and increases in evaluation demand, it is likely the SEU and OCB would stand to benefit from hiring additional manager(s). Further, the SEU may consider designating SEU focal points for key cross-cutting evaluation functions and stages, playing to the strengths of existing managers. For example, while all managers might be responsible for a portfolio of multiple open evaluations at any one time, one manager might be charged with ensuring sufficient evaluability processes have occurred across the unit and another may oversee ensuring use and follow-up are not going unaddressed. Further, it is evident that more work is needed to equip and empower evaluation commissioners and intended users at the cell-level to adequately carry forward evaluation findings and recommendations. While cost-utility analysis has been conducted, total and average evaluation budgets relative to total evaluation object program budgets have not been compared to arrive at any statements about the relative appropriateness of evaluation budgets to operational budgets. UNEG guidance suggests 3% - 5% of total program budget should be allocated to the evaluation function. Consider comparing average total expenses for OCB for the years of the evaluations with the total annual evaluation budget to see how the evaluation function budget relative to total expenses compares to this or other industry standard recommendations.

[43] SEU Roles and Responsibilities; Survey data about responsibility of roles; Interviews with Head of SEU; Interviews with SEU managers; Other SEU policy documents.

3. EVALUATION POLICY: VERY GOOD

Definition: "Every organization should establish an explicit evaluation policy. Taking into account the specificities of the organization's requirements, the evaluation policy should include a clear explanation of the purpose, concepts, rules and use of evaluation within the organization; the institutional framework and roles and responsibilities; measures to safeguard evaluation independence and public accountability; benchmarks for financing the evaluation function that are commensurate with the size and function of the organization; measures to ensure the quality and the use of evaluations and post-evaluation follow-up; a framework for decentralized evaluations, where applicable; and provision for periodic peer review or external assessment. The evaluation policy should be approved by the governing body and/ or the executive head to ensure it has a formally recognized status at the highest levels of the organization. References to evaluators in the policy should encompass staff of the evaluation function as well as evaluation consultants."

SEU Description: Strong evidence[44] demonstrates the SEU has "Very Good" *Evaluation Policy.* The SEU has a decent array of evaluation guidance documents that articulate and explain in detail the values, principles, and procedures that constitute good evaluation practice at OCB. Together, this evaluation policy contextualizes many industry standards to institutional and operational settings. Roles and responsibilities are delineated, as well as specific actions for each role by the SEU-specific evaluation stages. Participant observation and review of artifacts suggest these policies are being enacted for the most part consistently across evaluation cases. A systematic policy around external assessment or peer review (meta-evaluation) is absent. Additionally, at the outset of the meta-evaluation inquiry, the SEU indicated no quality framework was developed to serve as criteria and standards for the review and that one would need to be developed. The meta-evaluators were surprised to encounter the breadth of quality statements across the policy documents and wondered why these policies and statements were not foregrounded more fully as the basis for meta-evaluative claims.

Recommendation: The SEU can be more explicit in owning components of policy documents as the basis for quality assessment and meta-evaluation at OCB. Further, we suggest consolidating material across policy documents into one coherent policy document that houses values, principles, criteria, and standards of good evaluation practice. Further, any eventual consolidated quality framework might consider what principles or practices, either presently included or absent from SEU documents, are exclusive to the conduct of evaluating medical humanitarian interventions and need to be highlighted for external evaluation consultants to consider.

4. EVALUATION USE AND FOLLOW-UP: FAIR

Norm Definition: "Organizations should promote evaluation use and follow-up, using an interactive process that involves all stakeholders. Evaluation requires an explicit response by the governing authorities and/or management addressed by its recommendations that clearly states responsibilities and accountabilities. Management should integrate evaluation results and recommendations into its policies and programmes. The implementation of evaluation recommendations should be systematically followed up. A periodic report on the status of the implementation of the evaluation recommendations should be presented to the governing bodies and/or the head of the organization."

[44] SEU Guiding Documents; SEU annual goals.

SEU Description: There is strong evidence that Evaluation Use and Follow-up are "Fair" at OCB.[45] This rating is the result of splitting the difference between the two constructs the UNEG framework combines for this norm–Follow-up and Use. These should ideally be two distinct Norms that receive their own measures and ratings as an evaluation can be used without it being followed up, or followed up only to find out there was no use. With that, limited management responses and follow-up documentation resulted in a "Poor" Follow-up rating. Detailed analyses about evaluation use and use outcomes in the mEQ3 findings section reveal evaluation use at OCB is "Good."

Recommendations: The meta-evaluation team acknowledges mere application of evaluation recommendations is not the only or most important marker of quality in terms of use and follow-up. It may be that recommendations from external evaluators are poorly supported, not feasible, inappropriate, untethered to evaluative conclusions and findings, and not culturally responsive or specific. Differences of opinion remain about the role of evaluators in making recommendations and the reviewers appreciate the SEU's position about the collaborative nature of recommendation generation, per policy documents. With that, the MSF context does seem to place high value on "findings use" or being able to make decisions and act from quality recommendations that follow logically from sound evaluative conclusions. The value placed on this type of intended use of evaluation and the actual performance of this follow-up function suggest a major disconnect in aspiration and reality. Some significant remediation plan for evaluation follow-up is needed from the SEU and SEU steering committee. While it is the view of the meta-evaluators that the responsibility for evaluation use rests across many roles, the evaluation commissioner is primarily responsible for evaluation use and the SEU is primarily responsible for evaluation use follow up. Consider developing a recommendation rubric that defines the dimensions of a quality recommendation according to the SEU and OCB. Invite evaluators to use this rubric when creating or co-creating recommendations with evaluation participants. Invite consultation group members and especially commissioners to use this in their follow-up and response. Finally, consider having both commissioners draft a functional management response in collaboration with the SEU head of unit that is published as an annex of the meta-evaluation to increase transparency and credibility in terms of committing to areas that the meta-evaluation identifies as needing improvement. Write evaluation policy and enact procedures that duplicate this practice for primary evaluations, that is, include a management response as a public annex by default. Create a rubric of general evaluation quality in that management response template that managers can rate so the content of their responses is more transparent.

5. HUMAN RIGHTS AND GENDER EQUALITY: FAIR

Definition: "The universally recognized values and principles of human rights and gender equality need to be integrated into all stages of an evaluation. It is the responsibility of evaluators and evaluation managers to ensure that these values are respected, addressed, and promoted, underpinning the commitment to the principle of 'no-one left behind.'"

SEU Description: Strong evidence[46] suggests the SEU does a "Fair" job of attending to *Human Rights and Gender Equality* in evaluations. While the SEU does not espouse an explicitly human rights-based or -centered approach to evaluation, attention to basic human rights is attended to in the SEU ethical policy document, which is predominantly sourced from the UNEG Norms and Standards document.

[45] Evidence sources include: Use and Dissemination Plans; Survey response data on evaluation use and consequences; Management Response Documents; Use Satisfaction Ratings; Use and Influence Ratings.

[46] ALNAP 1.2, 2.5, 4.4 Gender Standard; PrgES P3; Evaluation Contracts; Evaluation Policies (framework, manifesto, roles and responsibilities)

While the SEU portfolio is rated “good” for the Propriety criterion, there are sub-criteria and sub-domains that are not as well rated. Only a third of the managed evaluations across the portfolio attended to gender in analysis, and about a third were missing cross-cutting analyses attending to vulnerable populations and protection, despite these being two strong values for MSF. Most evaluation cases did not produce evidence that groups traditionally excluded from or hindered by evaluation processes were sought out.

Recommendation: Recommendation: Consider referring to UNEG guidance and technical scorecard for ideas on mainstreaming human rights and gender equity in evaluation practice. Additionally, following comments with the head of unit about internal political will about MSF’s attention to institutional complicity with global structures of exploitation and colonization, following the “Do No Harm” humanitarian principle, commission an internal or external transversal desk report on transformative evaluation approaches, and specifically Culturally Responsive and Equitable Evaluation (CREE) models and evaluation policies and the potential benefit these practices provide OCB in being a leader in aligning practices with aspirations with their evaluation function. A possible entry-point to this conversation could be comparing the analysis of the Cultural Reading of the 2nd Edition of the Program Evaluation Standards with the PrgES framework that was used for this meta-evaluation. An additional source of resources might be the Funder and Evaluator Affinity Network Call to Action series that presents concrete issues related to equity from common evaluation policies and practices. There is also this related document: Righting Systemic Wrongs Organizational Self-Assessment.

6. IMPARTIALITY: FAIR

Norm Definition: “The key elements of impartiality are objectivity, professional integrity, and absence of bias. The requirement for impartiality exists at all stages of the evaluation process, including planning an evaluation, formulating the mandate and scope, selecting the evaluation team, providing access to stakeholders, conducting the evaluation, and formulating findings and recommendations. Evaluators need to be impartial, implying that evaluation team members must not have been (or expect to be in the near future) directly responsible for the policy setting, design, or management of the evaluation subject.”

SEU Description: Strong evidence[47] suggests the SEU does a “Fair” job of attending to issues of *Impartiality.* Though housed within the operations department, members of the SEU are not involved with the policy-setting, design, or management of the objects of evaluation, which is good. Many of the SEU managers were hired internally, which increases their credibility and capacity to understand the institutional context, but despite this organizational familiarity, there is no evidence to suggest their experience prevents them from impartial management. For external evaluators, there were multiple instances of evaluation consultants being hired from within the MSF talent pool, meaning evaluators who may have first and foremost been subject-matter experts to the specific evaluation object with institutional experience supported by some degree of sufficient evaluation know-how. Members of the MSF movement place high value on evaluator contextual knowledge of the distinguishing values, principles, and behavioral commitments along with systematic operational constraints. However, this positionality may present a limitation in impartiality, especially if undisclosed in final reports with no recognition of or plans to mitigate potential biases. The relevant ALNAP standard that pertained to this norm systematically received “poor” ratings due to little to no

[47] Sources include: PrgES U8, F3, Proprietary Ratings, A3, A2; ALNAP 1.2 Rating; Evaluation Contracts; Evaluation Policies (framework, manifesto, roles and responsibilities); Participant observation.

reports of the nature and make-up of the evaluations team nor how these compositions may or may not bias evaluation processes. Further, inconsistent attestations of conflicts of interest in inception reports and final reports also lowered the score for the respective PrgES sub-criterion. Low reliability ratings also influenced this rating in terms of mitigating biases. However, multiple conversations with the head of the SEU included reflexivity to mitigate any undue biasing of this meta-evaluation process with prematurely disclosed judgments or personal opinions.

Recommendations: mandate conflict of interest statements in the evaluation report template; mandate a section within the inception report methods section that invites evaluator teams to speak to the evaluator or team composition and how their lived experience may constrain or enable more accurate evaluative conclusions; ensure managers invite evaluation teams to speak to how they plan on ensuring reliability in the inception report; hold discussions with the SEU steering committee about the benefits and drawbacks of SEUs policy (tacit or explicit) toward hiring former or current MSF movement members; make any policy decisions transparent.

Overall UNEG Standards Rating: GOOD

The UNEG Standards and sub-standards are more detailed quality dimensions and sub-dimensions than the general evaluation Norms and assess the extent of the health of an evaluation function within an organization, including the conduct of evaluations. Against these standards, the SEU was awarded an overall "Good" rating. Many of these dimensions co-relate to quality dimensions in the other quality frameworks, as well as general Norms. When this occurs, the definitions and sub-standard often differ by nature and degree to some extent. This "Good" rating represents a healthy institutional framework, competent management function, and robust quality assurance measures for the evaluation function at OCB. Table 7 provides a summary of standard and substandard ratings, which is followed by key analyses of important sub-standards not already addressed in the UNEG Norms or PrgES and ALNAP key findings segments.

Table 7. UNEG Standards and Sub-standards Ratings

UNEG Standards Ratings for SEU Evaluation Portfolio and System	
Standard and Substandard	Rating
MANAGEMENT OF THE EVALUATION FUNCTION	VERY GOOD
Head of Evaluation	VERY GOOD
Evaluation Guidelines	VERY GOOD
Responsiveness of Evaluation Function	VERY GOOD
EVALUATION COMPETENCIES	VERY GOOD
Competencies	VERY GOOD
Ethics	GOOD
INSTITUTIONAL FRAMEWORK	GOOD
Institutional Framework for Evaluation	VERY GOOD
Evaluation Policy	VERY GOOD

Evaluation plan and reporting	VERY GOOD
Disclosure Policy	GOOD
Management response and Follow-up	POOR
CONDUCT OF EVALUATIONS	GOOD
Terms of Reference	VERY GOOD
Stakeholder Engagement and Reference Groups	VERY GOOD
Communication and Dissemination	VERY GOOD
Timeliness and Intentionality	GOOD
Scope and Objectives	GOOD
Methodology	GOOD
Selection and Composition of Evaluation Teams	GOOD
Evaluation Report and Products	GOOD
Recommendations	GOOD
Human Rights-based Approach and Gender Mainstreaming	FAIR
Evaluability Assessment	POOR
QUALITY	GOOD
Quality Assurance System	GOOD
Quality of the evaluation design	GOOD
Quality of the final stage of the evaluation	FAIR
OVERALL STANDARD RATING FOR EVAL PORTFOLIO and SYSTEM	GOOD

1. INSTITUTIONAL FRAMEWORK FOR EVALUATION: VERY GOOD

Definition: "The organization should have an adequate institutional framework for the effective management of its evaluation function."

SEU Description: all available evidence[48] suggests the SEU has a "Very Good" *Institutional Framework for Evaluation*. The SEU has an adequate support structure in terms of steering committee, support from board, executive leadership, adequate human resources. Data on use and influence suggest improvements could be made in extending integration of evaluation function with management decisions at cell and operational centre level. No formal analysis of SEU budget relative to total operational budget has been conducted, but such analysis would provide insight into the relative appropriateness of the SEU framework for OCB. Continued improvements in evaluation quality, and especially evaluation value (through improvements of the use and follow-up) will strengthen OCB

[48] SEU Evaluation Framework Document; Key informant interviews.

management's understanding and support for the evaluation function to contributing to the effectiveness of the operational centre.

Recommendation: Consider the head of unit analyzing percent of SEU budget relative to total operational budget and report to the steering committee and OCB board the degree of appropriateness of financial and human resource allocations. Ensure meta-evaluation conclusions and recommendations are weighed, prioritized, and translated into a use plan for evaluation system quality sustainment and improvements. The SEU SC should greenlight the SEU's inclusion of a new or updated policy around the ideal roles/responsibilities/expectations of 1) commissioners and 2) primary intended users (requesters) for ensuring evaluation use is maximized. We recognize roles and responsibilities of evaluators, managers, and commissioners are articulated, these may need to be updated and revisited with inclusion of primary intended users groups, if not already. This may entail the agreement to certain procedures lead by the SEU in facilitating more concerted evaluation follow-up. This may also include consultation with the SEU during scoping, inception, and throughout the life of the evaluation about possible intended uses for intended users based on possible evaluation outcomes. This may also include communicating more fully an active role of evaluators to take in facilitating intended use during evaluation beyond dissemination activities, which can be worked into existing touchpoints with CG and intended users.

2. COMMUNICATION AND DISSEMINATION: VERY GOOD

Definition: "Communication and dissemination are integral and essential parts of evaluations. Evaluation functions should have an effective strategy for communication and dissemination that is focused on enhancing evaluation use."

SEU Description: Strong evidence[49] reveals the SEU does "Very Good" at *Communication and Dissemination.* The SEU has a standard procedure of developing Use and Dissemination plans, with 71% (12/17) of evaluations having plans from the first instance of use in 2020. These are mostly communication and dissemination plans, but often do speak to use and utilization in terms of key meetings or decision points with specific activities to support operational management. The overall portfolio of evaluations for the 5 years saw 39% of evaluations with use and dissemination plans. Even in the absence of these plans, multiple reporting formats and reporting moments or events such as workshops, roundtables, and webinars have been a main feature. Where evaluations had specific Use and Dissemination Plans, many survey respondents indicated most of the activities were fulfilled. It is unclear the extent to which Use and Dissemination plans are used as management tools for follow-up.

Recommendation: Continue to develop and use Use and Dissemination plans and recommit to using these plans as the map for significant re-engagement with increasing actual evaluation use and follow-up. We strongly encourage the SEU to reclaim the authority to follow-up with recommendations as well as specific activities that facilitate use.

3. EVALUATION REPORT AND PRODUCTS: GOOD

Definition: "The final evaluation report should be logically structured and contain evidence-based findings, conclusions, and recommendations. The products emanating from evaluations should be designed to the needs of its intended users."

[49] Use and Dissemination Plans; PrgES P5 and A8 ratings; ALNAP Section 5.

SEU Description: strong evidence[50] reveal the SEU manages the delivery of "Good" *Evaluation Reports and Products.* Portfolio-wide scores for report quality using the ALNAP standards were high with an average of 88% and a Very Good Rating. These are minimum standards, but indicative of overall report quality, resulting from templated reports, and multiple quality assessment points. PrgES Accuracy ratings are lower, with lower scores for the following sub-criteria: explicit evaluative reasoning, information management, reliable information, and justified conclusions and decisions. Also, of importance and related to the logic of evaluative reasoning, most all reports had criteria or dimensions of merit (though not all) but many did not have explicit setting of standards, or degrees of quality or performance, for each criterion. This resulted in various practices for reporting actual performance against these standards from binary "met" or "not met", to meandering descriptions of mixed results with no conclusive judgment, to some using descriptors like "good" or "very good", but without the comparison of a qualitative scale and equally applied across criteria within the case. In short, most reports set out to ask "to what extent?" a performance criterion was meritorious, but did not establish standards of merit, either by the evaluator or collaboratively with evaluation users, to transparently answer this question. A positive note about evaluation reports and products is the effort the SEU takes to create multiple report formats for meeting different audience needs, from slide decks to full reports, to short versions, to posters, to reports in multiple languages.

Recommendation: Invite strengthening of evaluative reasoning by 1) underscoring existing guidance to connect findings, to questions and criteria; 2) link to evidence as much as possible such as in instances where findings were footnoted with evidence sources; 3) strengthen connection from findings to conclusions, where conclusions are not a discussion section of information not covered in findings; 4) ensure recommendations are tethered to specific conclusions or findings, and even potentially associated with criteria or evaluation questions. In short, overall strengthening connections of evaluative logic in reports can be improved, and likely met with ease with managers more fully articulating this value in templates and exchanges with consultants. A final and important recommendation we strongly suggest is to not stop at identifying criteria in the General Logic of Evaluation. Work with consultation groups and evaluators to establish standards of criteria, degrees of quality or goodness such as poor, fair, and good, and even adequate levels of quality or goodness for each individual criterion, depending on the nature of the evaluation object and evaluation purpose. The most intuitive place to integrate this practice is in an additional column in evaluation matrices in inception report annexes. Finally, the practice of multiple reporting formats should be continued and continually improved to meet audience needs. A light-touch desk review of optional evaluation reporting formats and potentially the extent to which different formats are preferred and have or could meet different MSF audience needs could make this strategy more efficient and effective.

4. RECOMMENDATIONS: GOOD

Definition: "Recommendations should be firmly based on evidence and analysis, clear, results-oriented and realistic in terms of implementation."

SEU Description: Evidence suggests[51] the SEU manages the delivery of evaluation reports with "Good" *Recommendations*. Recommendations were a feature of all but one evaluation with a total of 184 recommendations across 30 evaluations averaging 6 recommendations per evaluation. The PrgES does

[50] SEU Report template; ALNAP Section 5; Overall PrgES and ALNAP Ratings; Qualitative notes from reviewers.

[51] ALNAP Section 5.1.iii; Survey responses.

not have a standard for recommendations, likely resulting from a philosophical position about the necessity of recommendations in evaluations and more importance on evaluative conclusions. The ALNAP Proforma score for the recommendation standard was 94%, which is possibly inflated due to lack of deep contextual and organizational understanding from the report reviewers. Survey responses indicated recommendations are valued highly and likely the most scrutinized aspects of evaluations. Responses suggest recommendations in reports are often an unstated proxy indicator of quality including evaluator contextual understanding, which is also highly valued in OCB. This suggests OCB values instrumental evaluation use over conceptual and evaluation capacity development use.[52] Evaluators are equipped with adequate guidance in the evaluation report template and encouraged to make bold but warranted recommendations in their reports and more recently a practice of collaborative recommendation making has been encouraged through evaluation policy and report template guidance. One evaluation report indicated any changes between recommendations of the initial and final drafts of the final report. No transversal analysis was conducted to determine the relative quality of recommendations by case, but many survey respondents across roles noted variation in applicability, specificity, feasibility, and other considerations. These mixed qualitative ratings are weighed with more importance than our external document review and this rating reflects that.

Recommendation: Together with a special working group comprised of past consultation group members and intended evaluation users, the SEU might consider developing a recommendation matrix or rubric that formalizes the dimensions of recommendation quality and intended use, with such variables such as, but not limited to: ease of implementation; degree of anticipated benefit; program-specific or cross-cutting; feasibility; range of application; associated evaluation criterion, etc. Invite evaluators and consultation group members to assess and report assessments of recommendations according to these rubrics.

5. QUALITY CONTROL AT THE FINAL STAGE OF EVALUATION: FAIR

Definition: "Quality should be controlled during the final stage of evaluation."

SEU Description: Strong evidence[53] suggests end of evaluation quality control is "Fair." In-built feedback loops for the final stage of evaluation are the consultation group and manager comments on initial final drafts, a working session with the consultation group, and revisions to the final report. Additional reporting formats are developed by managers and the SEU coordinator to meet different audience needs. There is however a clear lack of management responses as well as recommendation follow-up, which the meta-evaluation team views as key factors in attesting to report quality, if not quality control in the case of management responses and ensuring use quality for recommendation follow-up. Also, while this meta-evaluation received detailed feedback about the final report from the primary intended user group–the SEU–feedback from the meta-evaluation consultation group was varied and inconsistent, suggesting the adoption of some standard feedback template or rubric for the CG and other user groups for primary evaluations would strengthen this quality dimension.

Recommendation: We strongly encourage a recommitment to meaningful management responses that are not seen as mere formalities, but as functional documents. We strongly encourage more resources dedicated from the SEU to follow up with evaluation use and consequences starting with a

[52] For more on evaluation use distinctions, see the section on use and influence under MEQ3.

[53] PrgES Evaluation Accountability Rating; KII; Management documents; Participant Observation.

recommitment to recommendation follow up procedures, as well as evaluation consequence inquiries possibly at 3, 6, and 12 months for each evaluation, depending on the nature and scope of the evaluation. Consider developing a user-friendly final report rubric for consultation group members to use in their reviews as a more systematic form of internal peer review. The SEU can pre-empt more useful feedback and possibly reduce the cognitive burden of consultation group members by providing a feedback template that uses a rubric of known dimensions of report quality along with open-ended questions. Evaluation commissioners can use the collection of CG feedback forms as inputs to their management response. Consider having the management response entail a final attestation to report/process quality in a rubric that is shared in the final report annex.

6. EVALUABILITY ASSESSMENT: POOR

Definition: "An assessment of evaluability should be undertaken as an initial step to increase the likelihood that an evaluation will provide timely and credible information for decision-making."

SEU Description: Strong evidence[54] suggests *Evaluability Assessment* practices at OCB are "Poor." Evaluation management artifacts such as pre-natal, stakeholder analysis document, evaluation checklists, and situational assessments suggest there have been attempts to determine and address evaluability. These limited documents were not shared with the meta-evaluation team, likely due to how infrequent they were used or that they were planned for but not developed. Scoping questions were shared for 8 of 31 evaluations. Interview data with the head of unit indicate significant insufficiency in terms of procedures related to what is known as evaluability assessments. Additional evidence such as survey data from evaluators, qualitative analysis of limitation sections, and actual performance of insufficient answers to scoped evaluation questions, mostly about outcomes and impact, suggest evaluability assessment is at best informal and inconsistent, and at worst non-existent. One formal evaluation policy, "Consider the evaluability of the project" exists as an operational principle under the "Method" domain in the evaluation manifesto. Seven of the 31 evaluations were awarded met standards for evaluability based on a generous interpretation of presence of scoping question documents. This standard is top two for most potential for improvement (the other being use and follow-up).

Recommendation: Identify which existing procedures can be modified to address more fully the evaluability question. Use existing or create an SEU-specific evaluability checklist to be integrated to the scoping stage of the evaluation process and conducted collaboratively with program teams. Identify go/no go standards or thresholds associated with each chosen dimension of evaluability, which may include some action in-between go/no go that augments the nature of the evaluation exercise if the evaluation is still deemed to be a net positive.

SEU EVALUATION MANIFESTO QUALITY FRAMEWORK RATINGS: VERY GOOD

As explained in the Evaluation Manifesto, this SEU framework is informed by the Program Evaluation Standards, the ALNAP Proforma, AEA evaluator competencies, and UNEG Norms and Standards. It is composed of various elements that register what quality means to the SEU, which are expressed as domains (broad categories), activities (evaluation processes), principles (prescriptive actions),

[54] SEU management documents; PrgES Feasibility Standard 2, checkpoint 1 ratings; ALNAP Section 1 and 4.3 ratings; KIIs.

products (documents), or events. A summary table of domains and sub-domains is reported followed by select key findings from each domain. These domains and sub-domains were drawn exclusively and without modification from the quality domains and sub-headings under the quality domains in the Evaluation Manifesto. Two approaches to valuing were used with this quality framework, qualitative ratings and re-coding PrgES and ALNAP indicators. All judgments were made by the external meta-evaluation team.

Overall Rating of EMQF: GOOD

When using SEU's official-unofficial quality evaluation framework (unmodified)[55], the SEU receives a "Good" rating. "Good" manifest adherence to this framework suggests the SEU is working hard to live up to its own standards of quality, no matter how fixed or unfixed. This framework was the first to see "Excellent" ratings for any domain or sub-domain, which were awarded to *Transversal Learning, Annual Report,* and *Evaluation Day*. The three poorest ratings (all "Poor") coincide with prior ratings of related domains in other frameworks and include, *Engage the voices of those less present, Follow up on findings and recommendations,* and *Consider the Evaluability of the Project.* This is a result of the EMQF being informed by these prior frameworks and the ratings being based on prior framework scores and ratings as well. This overall rating is a good baseline for the SEU that takes seriously its charge to deliver a quality evaluation function for OCB.

Table 8. EMQF Ratings for SEU Evaluation Portfolio and System

DOMAINS AND SUB-DOMAINS	RATING
VALUE (domain)	**GOOD**
Choosing Criteria (activity)	**VERY GOOD**
Ask the right questions (principle)	**GOOD**
Engagement and ownership (domain)	**GOOD**
Languages (domain)	**GOOD**
Ethics (domain)	**GOOD**
Engage the voices of those less present (principle)	**POOR**
USE (domain)	**GOOD**
Transversal Learning (domain)	**EXCELLENT**
Annual report (product)	**EXCELLENT**
Evaluation day (event)	**EXCELLENT**
External communication (domain)	**VERY GOOD**
Annual presentation at OCB board (event)	**VERY GOOD**

[55] The second valuing procedure in this meta-evaluation sub-question used a modified EMQF with an additional quality domain of *TRANSFORMATION.*

Real time learning (domain)	**GOOD**
Link to strategic platforms and meetings (principle)	**GOOD**
Communicate and disseminate findings (principle)	**GOOD**
Learning (domain)	**FAIR**
Follow up on findings and recommendations (principle)	**POOR**
METHOD (domain)	**GOOD**
Discuss Evaluator Competencies (principle)	**VERY GOOD**
Data (domain)	**GOOD**
Consider the Evaluability of the Project (principle)	**POOR**
OVERALL EMQF RATING FOR EVAL PORTFOLIO and SYSTEM	**GOOD**

ANNUAL REPORT (PRODUCT): EXCELLENT

SEU Description: evidence[56] from these reports demonstrate they exhibit "Excellent" potential for *Transversal Learning.* These reports that investigate the coverage of operational priorities and fidelity to strategic orientations are exceptional internal evaluative reports. This is one of the few instances of the SEU actually conducting their own evaluations, as opposed to managing them, and these limited examples are of high quality.

Recommendation: The reports investigate manifest fidelity to strategic orientations and operational priorities in evaluations conducted from the past year. The reports can be strengthened by adding elements of the closest existing model of evaluation, Principles-focused Evaluation[57], to these reports. Namely, in addition to fidelity to orientations and priorities, reports can investigate the relevance of orientations and priorities to objects of evaluations, and if relevant and adhered to, the results of adhering to orientations and priorities. Finally, consider including a section or emphasis of the report that makes similar meta-evaluative judgments about that year's evaluations to the relevance, fidelity, and impact of EMQF sub-domains reframed as effectiveness principles. For example, if the EMQF states "Engage the voices of those less present", investigate the extent of how relevant this was to past evaluations, if relevant, the extent to which it was adhered to in past evaluations, and if adhered to, what resulted from that adherence. This would suppose other elements of the EMQF currently framed as domains, products, or activities are reframed as effectiveness principles. For example, reframing "choosing criteria" to "choose the right and right number of criteria" provides more guidance than the current element name and could then be explained further as to what constitutes the right and right number of criteria. This reframing could theoretically be done for existing, and potentially additional, elements of the EMQF, assuming it is updated after the meta-evaluation.

[56] Annual reports.

[57] See the Annex XII on recommendation resources for more information.

EVALUATION DAY (EVENT): EXCELLENT

SEU Description: limited evidence[58] suggest this event is an "Excellent" investment in *Transversal Learning.* The evaluators were able to review some artifacts from evaluation days, including transversal analyses. This annual event signals an enabling evaluation environment, a commitment to promoting evaluation culture. From the outside looking in, evidence from these events suggest an excellent effort and performance for stimulating transversal learning.

Recommendation: Consider how to create lighter-touch events that re-energize OCB staff and promote a culture of evaluation. The meta-evaluation team is mindful of reports of webinar fatigue within OCB, so considering other creative activities is a possibility.

CHOOSING CRITERIA (ACTIVITY): VERY GOOD

SEU Description: strong evidence[59] suggests the SEU is "Very Good" at choosing, prescribing, and negotiating evaluation criteria. Across the portfolio, 97% (n=30/31) of evaluations used at least one OECD-DAC criterion, except for an evaluation that investigated an OCB budget overspend. The average evaluation used a combination of 3 OECD-DAC criteria. The overwhelming majority of evaluations the SEU manages are goal-oriented, with 87% (n=27/31) investigating effectiveness. Other regularly occurring OECD-DAC criteria were Relevance (68%, n=21/31), Efficiency (58%, n=18/31), Impact (52%, n=16/31), Sustainability (16%, n=5/31), and Coherence (6%, n=2/31). Among all evaluations, 94% (n=29/31) used additional evaluation criteria, with the most common criterion being "Appropriateness" occurring (51%, n=16/31) of the time. These findings suggest an appropriate balance of generally accepted criteria, and responsiveness to specific evaluation needs and contexts with custom criteria. There was evidence to suggest in the ALNAP 4.3 standard that some criteria were not evaluable based on several factors, but mostly data availability.

Recommendation: continue robust scoping practices with evaluation requestors, clients, and intended users that explore options for evaluation criteria. Integrate more evaluability assessment moments assessing data availability and purpose distinctions for evaluation scope. Consider continuing the practice of including fewer and higher quality assessed criteria, as opposed to all OECD-DAC criteria all the time. Review OECD-DAC guidance on the application of these criteria.

ENGAGEMENT AND OWNERSHIP (DOMAIN): GOOD

SEU Description: evidence[60] suggests that the SEU does "Good" at *Engagement and Ownership.* The portfolio of evaluations scores highest in Utility among all other criteria in the PrgES, which houses many sub-criteria about evaluation participant engagement. Propriety is another criterion that pertains to this, which received a "Good" rating. Given this element includes ownership, there are reports of less-than-ideal procedures around downward accountability with all right-to-know audiences. Further, there is evidence to suggest that evaluation users have low degrees of owning intended use plans.

Recommendations: revisit composition of consultation groups to include more perspectives from those who will be affected by the evaluation process. Re-assess use and dissemination plans to see

[58] Transversal analyses; SEU reporting; interviews.
[59] Evaluation portfolio review.
[60] PrgES Utility and Propriety ratings; surveys.

what feasible gains can be made in advancing this domain. Integrate transformative evaluation policies that address this domain through more culturally responsive and equitable evaluation practices.

DATA (DOMAIN): GOOD

SEU Description: evidence[61] suggest that SEU does a “Good” job of the management of evaluation data quality. A known and recurring issue within OCB is the lack of consistent monitoring data at the project level including the lack of clear program design theory that would dictate appropriate and needed types of data for management, let alone evaluation. However, evaluations regularly make use of existing data with 53% of all evaluations using secondary data analysis as a data collection and analysis method. Despite limitations, many evaluations used these existing routine health data to arrive at defensible and laudable data-informed evaluative judgments about effectiveness. Primary data collection methods were predominantly qualitative in nature with interviews, focus groups, and document reviews serving as primary data sources.

Recommendations: Integrate more robust evaluability assessment procedures into the scoping stages of evaluations. Lead out and demonstrate to operations cells/regional support teams what it looks like to plan for and collect data about programming by practicing this standard with your own programming for evaluation use and follow-up.

RE-CODING OF PRGES AND ALNAP INDICATORS TO MODIFIED EMQF DOMAINS: VERY GOOD (73%)

The second approach to valuing for the EMQF was to re-code the combined indicators of the PrgES and ALNAP frameworks according to a modified EMQF that included a fourth dimension of *Transformation.* Given the EMQF most closely aligns with a generally accepted conceptual framework about evaluation theory,[62] we decided to modify the EMQF to track with updated understandings about important issues of culturally responsive and equitable evaluation practice. Table 9 shows a summary of these reconfigured scores and ratings. What these results demonstrate is that using the more robust and well triangulated indicators of the PrgES and ALNAP frameworks, the SEU is doing “Very Good” for this modified EMQF. Our meta-evaluation team recommends serious consideration to the addition of this *Transformation* domain to the EMQF, in addition to any revisions and formalizations of this promising evaluation quality framework at OCB.

Table 9. Scores and Ratings of Recorded PrgES and ALNAP Indicators to Modified EMQF Domains

COMBINED PrgES and ALNAP					
Category	N	% of Total	Average Score	Rating	Rank
Use	38	17.04%	82.10%	Very Good	1
Methods	87	39.01%	76.06%	Very Good	2
Values	42	18.83%	71.54%	Very Good	3
Transformation	56	25.11%	63.16%	Good	4
Grand Total	223				

[61] Inception and final reports.

[62] See Annex XII for additional resources about the Evaluation Theory Tree.

MEQ3: WITH THESE META-EVALUATIVE CONCLUSIONS, TO WHAT EXTENT DO THESE COMPLETED EVALUATIONS PROVIDE VALUE TO OCB?

This meta-evaluation makes the distinction between quality and value. Quality is the merit of something. Value is the worth of something. Judgment claims of quality are based on absolute standards or ideal notions of what ought to be. Judgment claims of value are based on relative standards, or comparative notions of what else could be (often for the same costs). Both claims are dependent on the values of those making judgements. With that, the value of something is to some extent determined by the quality of that same thing. This study understands evaluation value as a function of evaluation inputs as budget costs, evaluation activity and output quality as Utility measures, evaluation immediate outcomes as evaluation use, and evaluation intermediate to long term outcomes consequences as evaluation use outcomes., and evaluation costs. The following analysis uses these premises to investigate the value of evaluations and the evaluation function at OCB. This is accomplished by 1) describing and judging the use and use outcomes of evaluations at OCB; 2) conducting a cost utility analysis that compares evaluation quality (in terms of utility) with evaluation cost (in terms of external evaluation consultant budgets); and 3) making an overall claim about the value of evaluations at OCB.

EVALUATION USE AND USE OUTCOMES

From the outset, meta-evaluation primary intended users at OCB expressed a special interest in identifying the extent to which evaluations were actually used as a function of the value prior evaluations provided to OCB. The meta-evaluation team did a poor job[63] of initially establishing shared definitions of evaluation use and consequences with meta-evaluation intended users. This resulted in the reliance of tacit and likely different understandings of evaluation use by the meta-evaluation team and primary intended users. When the initial final report draft was delivered, members of the SEU provided valid and useful feedback about the lack of a shared evaluation use definition and the potential inaccuracy of descriptions and judgments of evaluation use at OCB, notwithstanding limitations of little to no evaluation follow-up data. To address these specific concerns of the accuracy of these initial specific meta-evaluation claims about use, the meta-evaluation team revisited open-ended survey response data from evaluators, managers, commissioners, and project contacts for a second, more comprehensive analysis of evaluation use and outcomes at OCB. The following analysis presents a shared definition of evaluation use, updated descriptions of use and outcomes at OCB, and subsequent judgments about merit and value evaluation use and outcomes at OCB.

Definitions of Evaluation Use and Evaluation Use Outcomes

Framed as the "million-dollar question" by one primary intended user[64] on whether OCB evaluations were actually used, the same holds for transdisciplinary evaluation practice outside OCB. The results of evaluation processes and products is one of the most important issues in evaluation practice and has a long history of investigation and theorizing. Many thought-leaders have conceptualized the results of evaluation as: utilization, use, misuse, influence, consequences, impact, constitutive effects, and unintended effects. The generally accepted shorthand for the results of evaluation is *evaluation*

[63] See the meta-meta-evaluation attestation in Annex XIV.
[64] Key informant interview.

use. Though there are differences of understanding in the evaluation field,[65] at a high-level, the meta-evaluation team understands the results of evaluation practice as "the changes that do (or do not) occur as a result of evaluation."[66] These results include what is traditionally understood as evaluation use–how primary and secondary intended users think or act differently as a result of evaluation processes and products–and the outcomes of evaluation use–the results or the effects of those changes in thought or behavior among primary intended users. Desirable evaluation use and use outcomes are often a result of quality evaluation utility–the ways intended uses for intended users are attended to by evaluators and evaluation participants in evaluation design, planning, and execution.

Two approaches to understanding and observing evaluation use exist: 1) thinking of use as a broad sensitizing principle, or 2) understanding use more concretely as a specific typology of changes. There are benefits and drawbacks to both approaches.[67] For clarity and ease of measurement for this study, the meta-evaluation team used the *typology of change* approach to observing evaluation use at OCB. This typology[68] classifies use types by 1) the source of use, 2) purpose of use, 3) program aspect, 4) user type, 5) degree of influence, and 6) timing of use. For simplicity, the meta-evaluation team used a streamlined typology combining *use source*, *use purpose,* and *degree of influence* to describe use and use outcomes at OCB as shown in Figure 2. *[69]*

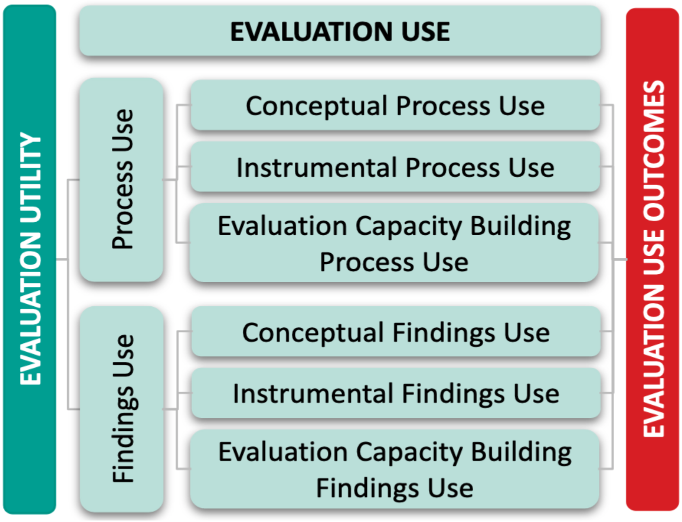


Figure 2. Evaluation Utility, Use Types and Use Outcomes

There are two main sources or stimuli of use in evaluations, evaluation processes and evaluation findings, respectively referred to as process use and findings use. *Process use* refers to how individuals and organizations make changes just by being involved in the evaluation process. *Findings use* refers to how individuals and organizations make changes by the delivery of an evaluation product, such as findings and or recommendations, typically included in some type of oral or written report. For each use source, there are generally five purposes of evaluation use. These include, 1) *conceptual use:* changes in understandings about the nature of the object of evaluation or the nature of evaluation itself; 2) *instrumental use:* changes in actions taken, typically resulting from decision making; 3)

[65] For the best overview of the last 50 years of discourse and research on evaluation use, see Alkin & King's 3-paper series and Patton's response, all of which are linked to in the additional resources annex.
[66] Mark, 2007, p. 117
[67] See Patton, 2020
[68] See Alkin & King, 2017
69 Not shown are legitimative use and symbolic use, along with other sub-dimensions such as user type, program aspect, timing, and degree of influence.

evaluation capacity building: changes in individuals or organizations ability to request, design, manage, execute, consume, and use evaluations; 4) *legitimative use:* commissioning evaluations to provide support or justification of decisions already made, and 5) *symbolic use:* commissioning evaluations as a political show, as public relations, or to placate stakeholders and delay meaningful action. Certain instances of legitimative use and all instances of symbolic use should really be viewed as evaluation misuse. Finally, the degree of use influence can be thought of as the significance or importance of the outcomes that result from these use types. Again, the meta-evaluation team distinguishes between how evaluation users think or act differently–evaluation use–and the positive, negative, intended, or unintended results of those changes–evaluation use outcomes. Just because a decision was taken or an understanding was updated, does not automatically mean the new idea or decision was beneficial and led to desirable results for those involved in and affected by evaluations. Evaluation use alone is a necessary, but insufficient proxy indication for estimating the value derived from evaluation processes and products.

Description of Evaluation Use and Outcomes at OCB

Fidelity to evaluation policies for evaluation follow-up is POOR as indicated by ratings for UNEG Norm 14 *Follow-up and Use*, EMQF sub-dimension *Follow-up on Findings and Recommendations*, and PrgES checkpoint 6 of *Utility* sub-criterion 8–*Concern for Consequences and Influence*.[70] This resulted in little to no existing programmatic monitoring or secondary data for the meta-evaluation team to use for descriptions and judgments about use and consequences of the evaluation cases. As a note, if the SEU expects improved practices around systematic data collection and program monitoring at the cell and project level for operations, then they need to live up to that expectation for their own "programming" by updating and committing to policies and standard operating procedures for systematic sustained data collection for evaluation finding and recommendation application, other types of evaluation uses, and ultimately evaluation use consequences during and after the lifecycle of primary evaluations. Initial descriptions of evaluation use at OCB in the first meta-evaluation written report draft relied too much on descriptions of evaluation follow-up and did not adequately distinguish between follow up and use.

Primary data collected from online surveys provided updated descriptions of evaluation case use types and use outcomes. 57 respondents completed the online survey for 31 cases with the average case being represented by two respondents. Three evaluation cases had no survey respondents, 12 cases had only one respondent, nine cases had two respondents, and 10 cases had three respondents. No cases had all four participant roles (commissioner, manager, evaluator, project contact) respond. Some instances of cases with multiple respondents were due to multiple evaluators or commissioners participating in the same evaluation case. Instances of multiple respondents per case provided opportunities to corroborate or provide different reports of evaluation use and use outcomes in the same case. It should be noted that *instrumental use* is directly observable with records or reports of changes made due to decisions taken. *Conceptual use* is not as easily observable but was inferred from reports about meaningful dialogue of key issues for operations and evaluations.

Evaluation Use at OCB

Out of the total evaluation cases, 71% (n=22/31) had one or more survey respondents report at least one type of evaluation use. Of the total respondents, 61% (n=35/57) reported some type of identifiable

[70] The ALNAP Proforma does not investigate evaluation follow-up, use, or consequences.

evaluation use type. There were 46 total reports of evaluation use types across the 22 cases for an average of 2 use reports per evaluation case. Some use reports were corroborations of the same instance of use within the same evaluation case. Other instances of multiple use reports within the same case were of different use types, such as one evaluation case having both conceptual findings use and instrumental findings use, or some other combination. Use reports that were stimulated by report findings were the most common with 52% of total reports being conceptual findings use (n=24/46) and 35% being instrumental findings use (n=16/46). Process use reports constituted only 7% (n=3/46) of reports with 3 instances of conceptual process use and 1 instance of instrumental process use. Combining both findings and process sources, conceptual use was the most frequently reported use purpose comprising 59% of all use reports (n=27/46), followed by instrumental use at 37% (n=17/46). There was one report of evaluation capacity building from evaluation findings in the *Treatment & Rehabilitation of Victims of Torture Programs* evaluation case, where findings were used to revise indicators for monitoring VoT programs. Finally, there was one report of legitimative use from findings with the *Reaction Assessment Collaboration Hub: Reach Project* evaluation, though the instance seemed neutral to positive and not an instance of misuse as it “Enable[ed] decision making that was already planned, but with solid evidence and also clearly identified value of product and process.”

Combinations of use types were common. The *Arche Project: Centre of Traumatology* evaluation had reports of both instrumental process use–when users made adaptations after debriefing with evaluators at the end of a site visit, and instrumental findings use–with the implementation of evaluation report recommendations. The *Eshowe HIV Project* evaluation had both types of conceptual use–process and findings–as indicated from a respondent sharing that, “During the evaluation *process* and while discussing the *results*, it [triggered] very interesting and honest conversations and reflections among MSF stakeholders about key aspects (such as community engagement, soft power...)” emphasis added. The *Corridor Programs for Key Populations* evaluation had both types of findings use, first conceptual findings use with the report that “Evaluation findings [have been] used for many years, as part of project discussions and planning, including with external stakeholders” and instrumental use that the evaluation contributed to a decision to continue specific investments within the community.

Evaluation Use Outcomes at OCB

While it is evident that evaluations are used at OCB what is less evident are the outcomes of those use reports. Nine cases, or 29% of the total 31 cases, or 41% of the cases that reported actual use (n=9/22) had reports of what could be understood as evaluation use outcomes–descriptions of what changes in understanding and behavior resulted in for those who use and are affected by evaluations. However, some readers could also interpret a few of these outcome descriptions as just additional instances of instrumental use. Eight out of the nine evaluation cases coded to have reports of evaluation use outcomes also had multiple use types reported. This suggests planning for multiple instances and types of use for evaluation cases may increase the probability of seeing use outcomes.

Some examples of use outcomes include the *Arche Project: Centre of Traumatology* evaluation, where instrumental process and findings use resulted in reports of positive adaptations of implementation and improved exit processes. Conceptual process and findings use types in the *COVID-19 Digital Health Promotion* evaluation case resulted in renewed investment from the medical department in human resources. The same combination of conceptual process and findings use in the *HIV Decentralization Initiative* evaluation case contributed to increased understanding among evaluation users about the nature and value of evaluation, which led to additional evaluation commissions. The manager shared that, "Conversations during this period helped MSF stakeholders to understand the potential added value of evaluations in general, and more concretely to identify the relevance of another evaluation focusing on currently implemented activities. This evaluation materialized some years after, and it's being currently implemented." Conceptual findings use in the *Hurricane Matthews Emergency Response* evaluation resulted in updated standard operating procedures for emergency response regarding shelter, which were deemed improvements by one of the respondents. Conceptual and instrumental findings use in the *Torture Rehabilitation Project* evaluation led to a report of "Better planning and MSF positioning in public advocacy regarding victims of torture."

There was one negative report of an evaluation use outcome by a respondent from the *OCB Operational Prospects 2014-2017 Review* evaluation that stated a "recommendation to reduce the importance of output indicator [was] a bad idea and has contributed to new problems." This highlights the premise that evidence of instrumental findings use, in this case following through on a recommendation, does not necessarily equate to value for intended users. This serves as a reminder for the SEU in ensuring updated evaluation follow-up procedures do not over-index on whether recommendations were applied or not as the only or main indicator of successful evaluation use.

Meta-evaluative claims about evaluation use and use outcomes at OCB

In the absence of generally accepted rubrics for judging the merit of evaluation use, there are a few logical approaches to making value claims about the state of use and use outcomes at OCB. A first approach might be to examine the distribution of use types by asking, "Are we seeing the types of use we value?" Initial meta-evaluation draft report feedback revealed the SEU values a broad combination of use types, not just the classic notion of instrumental findings use of applying report recommendations. Instrumental use may be valued mostly by those in the pragmatic MSF movement, but the SEU feedback rightly acknowledges the importance and potential longer lasting influence of conceptual use. What this study reveals is that there is a good amount of findings use at OCB, and within that source of use there is a good balance of conceptual and instrumental use purposes. Less present are reports of use stemming from processes and use for the purposes of evaluation capacity building. Changes from being involved in processes may be harder to observe, if not done close to the time of evaluation, and it may be that evaluation capacity building is a lower priority for OCB. The neutral to positive instance of legitimative use and lack of symbolic use or misuse is also a positive finding for OCB.

Another approach is asking whether there are outcomes of use and, if so, if those outcomes are benefiting intended users and those affected by evaluations. Roughly a third of total evaluation cases (n=9/31) had what might be considered evaluation use outcomes, with just 40% (n=9/22) of cases that reported use also reporting use outcomes. Further, the meta-evaluation team coded evaluation cases by their degree of use and use outcome as either low, medium, or high. These ratings were based on the extent to which use reports were corroborated, compelling, and inferred to be significant by the nature of use outcome descriptions. Of the 35 respondents who reported one or more instances of evaluation use, 60% of reports (n=21/35) were rated having "Low" degree of use and use influence. 34% of reports (n=12/35) were rated having "medium" degree of use and use influence. Two reports of use and influence were rated as "High." The first came from the manager of the *Corridor Programs for Key Populations* evaluation that reported both conceptual findings use had lasted years and instrumental findings use led to adequately addressing community needs. The second was with the

HIV Decentralization Initiative evaluation, which conceptual process and findings uses were reported to lead to further evaluation work. We did not compare these ratings of use and influence with the quality ratings of evaluation cases. It could be that unproductive instrumental use is based on poor quality findings (or recommendations as with the one instance from the *Operational Prospects 2014-2017 Review* evaluation case), or that productive conceptual use is based on poorer quality processes, or other combinations of evaluation quality ratings and use types and their consequences. Indeed, not captured in the survey data was an instance of the head of unit reporting conceptual use stemming from a less than ideal evaluation process.

Combining the evidence, descriptions, and reasoning with the generic and unqualified[71] qualitative standards rubric used throughout this meta-evaluation (Poor, Fair, Good, Very Good, and Excellent), the meta-evaluation team can provide provisional value claims that *evaluation use* at OCB is “GOOD” and that *evaluation use outcomes* are “FAIR.” Different meta-evaluation participants may arrive at different conclusions with the combination of additional first-hand or secondary evidence, rubrics, and reasoning. With the absence of regularly collected follow-up data that could have provided more accurate and timely descriptions of evaluation use and use outcomes, the descriptions and judgments herein may systematically under-estimate actual use and use outcomes.

COST ANALYSIS

The intent of this cost analysis is to understand the relationship between money spent for an evaluation budget and the relative utility gained from the investment. In the Figure 3 below, we chart the standardized PrgES Utility score and evaluation budgets to see how they fluctuated in relation to one another, asking *“what is the relative value of utility?”* Projects to the left of center scored lower than the mean on utility and those to the right, scored higher than the utility mean. Where these indicators are highly incongruent, we have a relative mismatch between budget and utility. For instance, the *End-to-End Supply Chain* evaluation held the highest spot on the budget z-score but was just above the mean on utility (a LOT was spent yet provided mediocre utility!). Useful, but perhaps less spectacular, are the projects that fell about the same position on both budget and utility. For instance, above the mean, *Maternal and Child Sexual and Reproductive Health Intervention*, *Adolescents Sexual and Reproductive Health Project*, *Clinical Mentoring in MSF's Non-Communicable Disease Project*, *Cervical Cancer Intervention*, and the *Catalytic Role of Mumbai Project with Regards to Policy Changes* evaluations all scored about the same on both measures. These could be considered good investments because, though they spent more money, they also produced stronger utility. The *Hospital Management Unity* and *Reaction Assessment Collaboration Hub: Reach Project* evaluation*,* on the other hand, scored very high on utility for much less money and the *Arche Project: Centre of Traumatology* evaluation was above the mean on utility, but below the mean on budget (good utility for less money). In total, there are 7 evaluations that score above the mean on utility but fell below the mean on budgets. These are evaluations that, for their cost, still produced higher than average utility.

We also tested the budget and the PrgES *Utility* score to see if they covaried. There was significant positive correlation between budget and PrgES *Utility* (r = .46, r2 = .21) with over 20% of all the variance in the model explaining the correlation.

[71] Without specific descriptions of what exactly constitutes each level, degree, or standard.

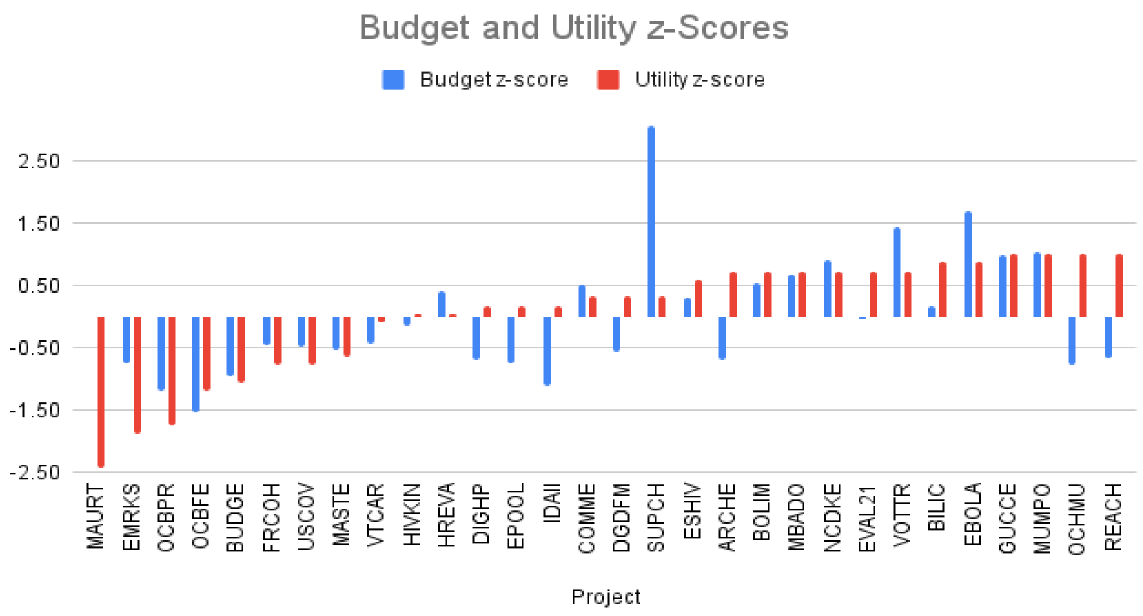


Figure 3. Budget and Utility z-Scores

QUALITY ADVANCEMENT

A correlation analysis (Spearman Rho) found a significant positive correlation between PrgES score and years, so that newer evaluations scored higher on the PrgES (r = .568, p = .001, r2 = .322). This does not hold for the ALNAP scores, which show a positive correlation, but it is very slight (r = .110, p = .554) and not significant. What this demonstrates is that the SEU is realizing actual gains in terms of evaluation quality and presumed evaluation value at OCB. Many meta-evaluation measures from this study can serve as a baseline for future portfolio meta-evaluations to more fully gauge progress against these standards. In the interim, comparative studies with other comparably sized and scoped internal evaluation units could be conducted with ALNAP scores and ratings for additional meta-evaluative insights.

VALUE PROPOSITION

From our analysis of the available evidence, we can conclude that OCB is getting good value, or worth, from its evaluation function. This judgment includes findings and conclusions about the state of evaluation use and use outcomes, the cost-utility of evaluation cases, and the change in evaluation quality scores over time at OCB. *Evaluation use* was rated “Good.” *Evaluation use outcomes* were rated “Fair.” Almost a quarter (n= 7/31) of evaluation cases in the last 5 years provide above the mean *Utility* for budgets below the mean; and only 10 of the 31 evaluations (33%) scored below the mean on *Utility*. Correlational analysis indicates that overall evaluation quality is increasing. These are encouraging conclusions that should be celebrated and serve as a baseline for further conversations about the value of evaluation at OCB.

MEQ4: WHAT FACTORS WITHIN MSF ORGANIZATIONAL SPHERE OF INFLUENCE MEDIATE THE QUALITY OF EVALUATION PROCESSES AND PRODUCTS AND HOW CAN MSF USE THIS INFORMATION TO ENSURE HIGH QUALITY AND VALUE EVALUATIONS?

With an eye for intended instrumental findings use of this meta-evaluation, the evaluation team and intended users selected this final auxiliary question to go beyond defining, describing, and judging quality with an attempt to *explain* quality. The assumption here is that asking *why* evaluations were of high or low quality could lead to answers for *how* decision-makers could sustain and improve evaluation quality at OCB. While this study did uncover some potentially useful findings for this meta-evaluation question, the meta-evaluation team believes the SEU could develop an institutional learning agenda for evaluation improvement supported by ongoing light-touch internal inquiries or exploratory and explanatory exercises led by team members of the SEU that investigate the myriad of factors contributing to evaluation quality at OCB. Due to the explanatory, rather than evaluative, nature of this learning agenda, it may be that internal champions of evaluation within the SEU, SEU steering committee, and OCB–who have deep institutional knowledge and familiarity with organizational dynamics and culture–are better positioned to pursue such an agenda that could result in useful findings over short-term external consultants.

Five light-touch analytical procedures were applied for this question that included: 1) identifying systematic gaps in evidence and quality ratings; 2) coding PrgES indicators by the six-step evaluation process; 3) a maximum deviation analysis of the highest and lowest quality evaluation cases; 4) a transversal analysis of reported limitations from the 31 evaluation cases and; 5) exploring the responsibility of ensuring evaluation quality across the four key roles of evaluators, managers, commissioners, and project contacts.

QUALITY GAP ANALYSIS

The simplest analytical approach to answering what factors mediate quality–or what contributed to the quality scores and ratings we observed from the portfolio for evaluation performance criteria–is to investigate the ratings for the sub criteria and indicators. In other words, what were the explanatory factors for low or high quality–the presence or absence of specific indications of quality. Due to the transparent and systematic use of checklists, the SEU has a host of specific indicators across the PrgES and ALNAP frameworks that function as explanatory variables for low or high scores and ratings.[72] Investigating these is the most accessible answer to this question.

Exploring these factors can be done in two ways, from an asset-based approach for sustaining quality and or from a deficit-based approach for improving quality. Given the OCB performed well across so many dimensions of quality, a more feasible entry-point is to start with deficits, specifically the extreme and systematic gaps in quality. There are two reasons framework indicators could have received a “not met” rating.

[72] The UNEG Norms and Standards has what can be termed indicators of sub-criteria and criteria, but only portfolio-level ratings were applied to sub-criteria and criteria. The EMQF has descriptions of dimensions that could be operationalized into indicators for monitoring or measurement, as needed.

The first reason is due to systematic gaps in data availability, or *data gaps*. These are sub-criteria and indicators, typically process-oriented as opposed to product-oriented, that had no evidence or insufficient evidence to demonstrate standards were regularly met across the portfolio. For propriety and accuracy considerations, indicators with these data gaps were treated the same as indicators that had data that standards were not met.

These gaps could be the result of process-oriented standards being met, but not documented and not ultimately recalled by evaluation managers, evaluators, or the result of non-responses from both roles (which happened in 3 evaluation cases). Or they could result from process-oriented standards not being met and not being documented they were not met. Some examples of the data gaps at the OCB we observed in this study include but are not limited to indicators for *Fiscal Responsibility, Information Management, Reliable Information, and Responsive and Inclusive Orientation.* Again, these data gaps could, but not necessarily, also be actual quality gaps. Systematic indicator data gaps were given special ratings of "not met*" for ease of distinguishing with the next category.

The second type of gaps are directly observable systematic quality gaps. These are sub-criteria and indicators that had sufficient evidence to demonstrate standards regularly not being met across the portfolio. No one evaluation case was perfect, and when viewing the METAE Dashboard, reading ratings for indicators across rows reveals where specific cases did or did not meet standards. Reading the Dashboard vertically provides insights into trends about standards for the portfolio. Using a cut-score of 25% for the PrgES framework, we identified indicators that were systematically unmet for an initial quality gap analysis. Twenty-three indicators, or 13% (n=23/180) fell into this category and are presented in Table 10 as priority quality gaps that the SEU and OCB can use for any follow-up planning.

Table 10. Quality Gap list Represented by PrgES Indicators with Scores of 25% or Less

The utility standards are intended to ensure that an evaluation is aligned with stakeholders' needs such that process uses, findings uses, and other appropriate influences are possible.	
Evaluator Credibility: Give stakeholders information on the evaluation plan's technical quality and practicality, e.g., as assessed by an independent evaluation expert.	**0%**
Attention to Stakeholders: Search out & invite input from groups or communities whose perspectives are typically excluded, especially stakeholders who might be hindered by the evaluation	**10%**
Concern for Consequences and Influence: Follow up evaluation reports to determine if and how stakeholders applied the findings	**10%**
The feasibility standards are intended to ensure that an evaluation is viable, realistic, contextually sensitive, responsive, prudent, diplomatic, politically viable, efficient, and cost effective.	
Practical Procedures: Assess and confirm the program's evaluability before deciding to proceed with the evaluation	**23%**
Resource Use: Document the evaluation's benefits, including contributions to program improvement, future funding, better informed stakeholders, and dissemination of effective services.	**0%**

The propriety standards are intended to ensure that an evaluation will be conducted properly, fairly, legally, ethically, and justly with respect to (1) evaluators' and stakeholders' ethical rights, responsibilities, and duties; (2) systems of relevant laws, regulations, and rules; and (3) roles and duties of professional evaluators.	
Clarity and Fairness: Develop and communicate rules that assure fairness and transparency in deciding how best to allocate available evaluation resources to address the possible competing needs of different evaluation stakeholders.	0%
The accuracy standards are intended to ensure that an evaluation employs sound theory, designs, methods, and reasoning in order to minimize inconsistencies, distortions, and misconceptions and produce and report truthful evaluation findings and conclusions.	
Justified Conclusions and Decisions: Identify the persons who determined the evaluation's conclusions, e.g., the evaluator using the obtained information plus inputs from a broad range of stakeholders	0%
Justified Conclusions and Decisions: Report plausible alternative explanations of the findings and explain why rival explanations were rejected	0%
Reliable Information: Determine, justify, and report the needed types of reliability—e/g., test-retest, findings from parallel groups, or ratings by multiple observers—and the acceptable levels of reliability	16%
Reliable Information: In the process of examining, strengthening, and reporting reliability, account for situations where assessments are or may be differentially reliable due to varying characteristics of persons and groups in the evaluation's context	23%
Reliable Information: Examine and discuss the consistency of scoring, categorization, and coding and between different sets of information, e.g., assessments by different observers	6%
Information Management: Document and maintain both the original and processed versions of obtained information	13%
Information Management: Retain the original and analyzed forms of information as long as authorized users need it	3%
Information Management: Store the evaluative information in ways that prevent direct and indirect alterations, distortions, destruction, or decay	19%
Sound Designs and Analyses: Plan specific procedures to avert and check for threats to reaching defensible conclusions, including analysis of factors of contextual complexity, examination of the sufficiency and validity of obtained information, checking on the plausibility of assumptions underlying the evaluation design, and assessment of the plausibility of alternative interpretations and conclusions	23%
Explicit Evaluation Reasoning: Examine and report how the evaluation's judgments and conclusions are or are not consistent with the possibly varying value orientations and positions of different stakeholders	3%
Explicit Evaluation Reasoning: Identify, evaluate, and report the relative defensibility of alternative conclusions that might have been reached based on the obtained evidence	3%

Explicit Evaluation Reasoning: Assess and acknowledge limitations of the reasoning that led to the evaluation’s judgments and conclusions	**23%**
The evaluation accountability standards are intended to ensure that an evaluation is systematically, thoroughly, and transparently documented and then assessed, both internally and externally for its utility, feasibility, propriety, and accuracy.	
Evaluation Documentation: Evidence of the evaluation’s consequences, including stakeholders’ uses of findings	**3%**
Internal Meta-evaluation: Maintain and make available for inspection a record of all internal meta-evaluation steps, information, analyses, costs, and observed uses of the meta-evaluation findings	**0%**
Internal Meta-evaluation: Reach, justify, and report Judgments of the evaluation’s adherence to all of the meta-evaluation standards	**0%**
Internal Meta-evaluation: Make the internal meta-evaluation findings available to all authorized users	**0%**

SIX-STEP EVALUATION ANALYSIS

The next procedure to answer this question was to use the Six-Step Evaluation process as a diagnostic tool. Each of the 180 indicators for the PrgES were coded by the SEU six-step evaluation process and corresponding scores and ratings for the 31 evaluation cases were sorted by the new index of PrgES indicators by the six-step process codes. Scores and ratings for each step were derived as a diagnostic exercise to determine which steps may need improvement. The portfolio scored highest on *Inception* and tightly grouped were *Reporting*, *Data Collection and Analysis*, *Preparatory*, and *Scoping*, followed a bit further by *Dissemination and Use.* These findings suggest that the indicators for remediation associated with *Dissemination and Use* should be prioritized and may yield the most benefit in any post-meta-evaluation action plan. Next in line would be indicators that pertain to scoping activities.

Table 11. Six-step evaluation scores

STAGE	COUNT OF STANDARD	PORTFOLIO STAGE SCORE	PORTFOLIO STAGE RATING	RANK
Scoping	13	58.81%	Good	5
Preparatory	36	58.99%	Good	4
Inception	33	72.53%	Very Good	1
Data Collection and Analysis	45	60.65%	Good	3
Reporting	45	61.86%	Good	2
Dissemination and Use	8	44.35%	Good	6

MAXIMUM DEVIATION ANALYSIS FOR OUTCOME FACTORS

One of the questions posed to respondents of the meta-evaluation surveys was “What were the most important factors that determined the outcome of this evaluation?” In analyzing survey responses from both the highest and lowest scoring evaluation, two clear trends seemed to dominate respondents' views of what determined a positive or negative outcome. First, was the competency of the contracted external evaluation consultant(s). This included considerations of subject matter-expertise, responsiveness to stakeholders, and being a clear communicator. The presence or absence of these attributes in evaluators was indicated as a clear factor towards the evaluation outcome. There are at least two initial suggestions for how the SEU can use this finding. First, they can continue to use and refine the competency-based evaluator selection process in the *Preparatory* step. This includes refining the type of information included in the terms of reference and ensuring evaluations managers identify the most important evaluator competencies given the needs of primary intended users and the scope of the evaluation identified in the *Scoping* step and hiring for those needs. Second, the SEU update important internal quality checks or internal meta-evaluation moments or procedures in the inception phase, conveying additional important expectations about evaluation practice not conveyed in the terms of reference through updated content in the inception report template and the the use of an inception report review rubrics for the evaluator, manager, and consultation group to use.

The second most frequently reported factor said to determine the outcome of evaluations was evaluation participant engagement. This included high levels of engagement and collaboration from consultation groups, commissioners, and other participants. The presence or absence of this engagement was indicated as a clear factor towards the evaluation outcome. The SEU can consider identifying the promising practices managers already do to facilitate quality evaluation participant engagement at each step of the evaluation process, and then inviting all managers to adopt some or all of those practices. Utility is already a high scoring criterion, so some of these practices may be fine-tuning. Of note, consideration should be given to how patients and communities, especially those traditionally not involved in evaluations, are engaged throughout the evaluation process. Finally, as mentioned elsewhere, the SEU, SEU Steering Committee, and those from the Operations Department such as evaluation commissioners, clients, and primary intended user groups should co-create an appropriate strategy for maintaining participant engagement after the external evaluators have left the picture.

LIMITATION ANALYSIS

Each evaluation report reviewed included a limitations section detailing potential issues that might influence the accuracy, feasibility, or other dimensions of quality for the given evaluation. While many of the limitations were specific to discrete evaluations, a transversal analysis identified eight limitations that occurred frequently enough to suggest the evaluation system or environment could play a role in addressing these. While specific suggestions are offered for addressing these. An overarching approach to these known frequently occurring limitations may be to include them as points about the evaluation context that evaluators and participants should address in discussions and reports in the inception step. It should be noted that these limitations were reports from evaluators and a complementary but different set of limitations may have been reported by evaluation managers. It could be that an evaluation manager limitation section could be included in an updated management response or final report assessment rubric, which aggregated over time could serve as a source for transversal learning internally.

Short Evaluation Timelines: 41% of evaluations (n= 13/31) reported prescribed timelines were too short. Indeed, this meta-evaluation faced limitations in adequately scoping, forecasting, and executing all activities within the prescribed timeline. While there are elements of evaluation consultant management that may be outside the control of the SEU, what this finding might suggest is that many of the evaluations could be overscoped or suffer from mid-evaluation scope creep. Reducing the quantity of criteria applied, or evaluation questions asked may be an initial response to this issue. Reports from other evaluation participants in key informant interviews indicate some processes take longer than expected and should go faster, which may also suggest a mismatch in expectations and scope between participants.

Lack of Project Documentation: 35% of evaluations (n= 11/31) reported an absence of program or project documentation that limited how fast and how far the evaluation could go. Reports from key informant interviews revealed that instances where this limitation was present translated into additional time for evaluators to make sense of and just describe the object of evaluation, let alone pursue evaluative reasoning. While these descriptive activities were deemed valuable to the participants, these can still present threats to the feasibility criterion, which could affect others. This is squarely an evaluability assessment consideration. The SEU should recognize that deficits in organizational culture for adequate program documentation won't change overnight. In the short-term, measures should be taken to identify these risks through updated evaluability assessment activities in the Scoping step. Some thresholds of minimum viable evaluability standards should be established where go/no go decisions for commissioning evaluations. In the medium-term, mitigation strategies for addressing these known limitations before evaluation activities should be considered. In the long-term, OCB should consider with the SEU and SEU steering committee efforts for organizational evaluation capacity development, which would include improving program design and documentation among operational staff at the country, cell, and project level and may factor into the existing re-centralization strategy.

Lack of Monitoring Data: somewhat related, 35% of evaluations (n=11/31) reported a lack of monitoring data. Gains in the project documentation limitation will have a direct effect on this limitation. Without adequate program documentation, operations staff would not be able to adequately monitor performance data. The SEU should include this factor into evaluability assessments and scoping evaluations. The lack of documentation nor monitoring data should not necessarily be an automatic dealbreaker for commissioning evaluations, but it could scrub or stall evaluation commissions depending on the scale, scope, and nature of the evaluation. As mentioned elsewhere, the SEU should be a leader in modeling what it looks like to have healthy monitoring systems by revisiting mid and post-evaluation monitoring and follow-up. This is another evaluation capacity building domain that OCB could consider investing in at the operations level that starts with basic project and program design and management competencies.

Low availability of evaluation participants: 32% of evaluations (n= 10/31) reported low availability of evaluation participants negatively affecting the evaluation process and product. While there are likely instances of unplanned unavailability, this is a known dynamic of the highly demanding operational evaluation setting of OCB and adequate evaluation policies and procedures should be revisited to address this issue. Level-setting with commissioners, clients, and evaluators during scoping, preparatory, and inception steps may be all that is needed to address this issue. The SEU could flag this as early and ask evaluators to discuss ways to accommodate this feature of the context in their inception reports. In the mid to long-term the SEU can work with the SC and operations department might consider policies and activities that help prospective evaluation participants make more time for evaluation engagement at different steps.

COVID-19 Travel Restrictions: 26% of evaluations (n=8/31) reported travel restriction issues affected the evaluation. Despite gains in many national-level COVID responses where MSF operates, travel related restrictions may continue to be a limitation in future evaluations. The SEU may consider a light-touch internal desk review synthesizing the multitude of guidance resources for remote, distance-based, and tech-enabled evaluation practice that can be used as guidance for scoping, preparatory, and inception steps.

Data Quality Issues: 23% of evaluations (n= 7/31) reported issues with data quality. These related to evaluation informants and data sources providing inaccurate or incomplete data. The PrgES framework has multiple relevant indicators about reliable information that could be considered when addressing this factor of quality. In the short term, the SEU can place the burden of addressing this limitation on evaluation consultants by merely flagging it as a known issue in past evaluation contexts. In the mid-term, the SEU can plan for professional development activities that pertain to data quality, bias reduction, threats to validity, and reliability measures in data collection, as well as accommodating for political factors in evaluation data collection that may lead to data quality issues.

High personnel turnover: 19% of evaluations (n= 6/31) reported high personnel turnover affecting the evaluation process. Though somewhat related to the low availability of participant limitation this issue focused on staff participants and excludes patients and communities and other participants. This may have implications for data collection, but also evaluation management, and primary intended users at the cell level. Improvements in project documentations and monitoring data could mitigate the effects of this limitation for data collection, but internal transitions in the SEU and among primary intended users would not be addressed with those other domains. Ensuring there are multiple primary intended users from the client end involved through the six-step process could address this.

Recall Bias: 16% of evaluations (n= 5/31) reported issues with recall bias among informants and respondents. Like many of these limitations, this issue is not limited to MSF, but a feature of summative evaluations that cover years' worth of programming, including this meta-evaluation study as indicated in some survey responses. As with others, gains in program documentation and monitoring data can sufficiently address this issue, along with improvements in instrumentation of surveys or interview protocols that ask appropriate questions about past performance.

ROLE AND RESPONSIBILITY ANALYSIS

OCB staff who participated in the online survey were all asked, "How are [their role] responsible for the success or failure of evaluations at OCB?" The following segment presents the top themes from the responses of evaluation managers, commissioners (n= 9/39), and project contacts (n= 6/39). Evaluation managers (n= 24/39) had consensus that they had greater influence in the scoping, preparatory, and inception steps, and an important, but lesser degree of influence in data collection, reporting, and use and dissemination.

Managers most frequently reported responsibilities associated with planning for intended use (n= 10), followed by aligning expectations (n= 8), maintaining regular communication with and between evaluation teams and intended users (n= 7), general management activities for adaptive accompaniment of the evaluation (n= 7), and adequately identifying and mitigating risks to the evaluation (n= 6). Commissioners reported their most important responsibilities were to promote engagement among participants (n= 5), align expectations (n= 4), and apply findings (n= 2). Project contacts responses were limited, but respondents identified responsibilities for aligning expectations (n= 1), helping evaluators access information (n= 1), supporting the coordination for data collection or field visits (n= 1), guide the process with local knowledge (n= 1), and applying findings (n= 1).

Importantly, all three roles reported the shared responsibility of aligning expectations around evaluation scope, design, purpose, and use. Managers viewed their role as facilitator, coordinator, and safeguard in terms of quality and risk management. Commissioners recognized the legitimacy they provide to the evaluation process for operational staff. Project contacts viewed their roles as granting access and contextual knowledge. Both commissioners and project contacts shared an expectation that their roles should take responsibility for applying findings. These findings reveal important distinctions between these roles and how they can support one another and the evaluation process, as well as signal a shared degree of accountability for ensuring the quality of evaluations are high and that their potential value is realized with the alignment, planning, and application of findings for intended use.

>-<

CONCLUSIONS

DEFINITIONS OF EVALUATION QUALITY AT OCB ARE EMERGING AND DEFENSIBLE.

The SEU and OCB have a mature view of the nature of evaluation and an emerging quality framework informed by generally accepted evaluation quality frameworks. The collection of statements across SEU evaluation policy documents, responses from key informants, and the hybrid meta-evaluation framework constructed for this project provide a detailed picture of "What is Evaluation Quality at OCB?" for this study. It is expected that the SEU and SEU steering committee will continue to refine the answer to this question with updates to and consolidation of evaluation policy following this study.

EVALUATION QUALITY AT OCB IS GOOD TO VERY GOOD.

Using the definition of quality identified from this study, the portfolio of 31 evaluations managed by the SEU from 2017-2021 has been judged "Good" to "Very Good". The portfolio was judged "Good" using the Program Evaluation Standards, "Very Good" using the ALNAP Proforma, "Good" using the United Nations Evaluation Group Norms and Standards, and "Very Good" using modified and unmodified Evaluation Manifesto Quality Framework assessments. In concert, these ratings and historical analysis reveal evaluation quality at OCB is high and has improved over time.

EVALUATION USE AT OCB IS GOOD; USE OUTCOMES ARE FAIR; THE FULL EXTENT OF EVALUATION USE AND OUTCOMES IS STILL UNKNOWN.

Comparing the "Very Good" evaluation *Utility* rating with the "Good" *Evaluation Use* and "Fair" *Evaluation Use Outcome* ratings suggest a few conclusions. First, evaluators and managers are fulfilling their responsibility to prepare the conditions for evaluation use. Second, commissioners and clients within cell offices and operations departments can do more to make use of evaluation. Third, lower ratings of use, use outcomes, and degree of use and influence may be under-representative due to lacking SEU evaluation follow-up procedures and limitations with primary data collection and analysis measures of *Evaluation Use* from this meta-evaluation study.

OCB IS RECEIVING GOOD VALUE FROM THE EVALUATION FUNCTION.

Factoring utility, use, use outcomes, and evaluation costs, the portfolio of evaluations have provided good value or worth to OCB. The majority of most evaluations are used, used in multiple ways, and among multiple users. The best available evidence suggests use leads to positive outcomes for those involved and affected by evaluations, though benefits can be extended through concerted follow-up by managers and intended users. Cost Utility Analysis reveals good utilization for money.

THE EVALUATION SYSTEM AT OCB IS WELL FUNCTIONING AND HEALTHY.

Quality and value ratings for the 2017-2021 portfolio of evaluations is a barometer for the health of the evaluation system in which they were commissioned, managed, and used. Findings for evaluation quality and value indicate there is an enabling environment for useful evaluations at OCB. This is the result of the intentional and concerted efforts by the head of the SEU, evaluation managers, qualified external evaluation consultants, invested consultation groups, supportive evaluation commissioners, and helpful project contacts. It is also a likely result of a long-standing ambition to foster a culture of evaluation at OCB.

THE OCB HAS A ROADMAP FOR SUSTAINING AND IMPROVING QUALITY AND VALUE.

The chosen methods of meta-evaluation checklists and rubrics have resulted in transparent and actionable findings and recommendations for the SEU to sustain and improve evaluation quality and value at OCB. Strong quality dimensions to sustain include but are not limited to, *Utility, Reporting and Communication, Management of the Evaluation Function,* and *Transversal Learning.* Weak quality dimensions to improve include but are not limited to, *Accuracy*, *Evaluability Assessment, Explicit Evaluation Reasoning, Human Rights and Gender Equity,* and *Engage the Voices of Those Less Present.* Significant factors of evaluation quality identified from this study were evaluation competencies and evaluation participant engagement. Frequently occurring limitations were short evaluation timelines and lack of program documentation and monitoring data, among others. These findings and more have been accompanied with over 100 operational recommendations and five strategic recommendations to facilitate conceptual, instrumental, and evaluation capacity development evaluation use.

RECOMMENDATIONS

Influenced by a much-appreciated practice observed in one of the final reports from the evaluation portfolio, this meta-evaluation offered micro-recommendations that flowed logically from key evaluation findings. These operational recommendations are included in the findings section, as well as in annexes VI and VII. Following SEU guidance and tradition, the meta-evaluation team offers five main strategic recommendations deemed to have the most potential in assisting the SEU and OCB realize the value of this meta-evaluation through intentional and concerted follow-up.

⇒ Recommendation 1: **STRENGTHEN THE EVALUABILITY ASSESSMENT FUNCTION**

The SEU can realize significant gains in evaluation quality if existing scoping procedures were applied consistently and strengthened with the inclusion of evaluability assessment methods. These evaluability assessment methods can identify long-term and short-term areas for improvement to the institutional evaluation framework and overarching evaluation culture as well as identify important mitigation strategies for discrete evaluation cases, increasing greater resource use and potential evaluation use.

⇒ Recommendation 2: **RE-INVEST IN DOCUMENTING EVALUATION USE AND INFLUENCE**

If the SEU expects improvements to the evaluability of medical humanitarian interventions with at least the existence of consistent operations project monitoring of activity, output, and outcome-level data, then the SEU needs to lead out and model that expectation in that equivalent process of the monitoring of evaluation activity, output, and outcome-level data. Most importantly is the recommitment to policies for Evaluation Use Follow up through updated and consistent application of standard operating procedures for Evaluation Use Follow up. The SEU steering Committee needs to expect a more active role from evaluation commissioners and evaluation requesters in considering and committing to intended uses for intended users.

⇒ Recommendation 3: **DEMAND STRONGER EVALUATIVE LOGIC, REASONING, AND VALUING**

The SEU has a mature and healthy view of the nature and purpose of evaluation and its distinction from research and other inquiry modes to provide unique value to the management function of a high-stakes organization like MSF. The SEU needs to codify stronger expectations for evaluation consultants to be more explicit and transparent in their evaluative reasoning. This starts with the establishment of criteria standards for each evaluation criteria, identification of acceptable standard levels for each criteria, and the mock-up of intended instrumental uses if certain standards were observed during and at the end of evaluation studies. These can be facilitated through more frequent and intentional use of evaluation rubrics, which improve the transparency and credibility of evaluative judgments.

Recommendations 1-3 (of 5) →

⇒ Recommendation 4: **FORMALIZE THE INTERNAL META-EVALUATION FUNCTION**

The SEU has an assortment of quality assurance assets within the existing evaluation institutional framework. Taking cues from the previous recommendations, members of the SEU can improve the evaluative reasoning of external evaluation consultants by practicing improved meta-evaluative reasoning themselves through the formalization of the internal meta-evaluation function. This starts with drawing a more confident line in the sand about what constitutes evaluation quality, ideally through an updated and consolidated EMQF expressed as effectiveness principles. Additionally, the SEU should identify and articulate how sub-sector specific contingencies for medical humanitarian evaluation necessitate specific quality standards not presently found in the four existing quality frameworks. Meta-evaluation findings and recommendations should be used to realize that enhanced vision of quality. Finally, the integration of checklists and rubrics for key meta-evaluative moments would strengthen practical procedures for internal meta-evaluation.

⇒ Recommendation 5: **ADOPT TRANSFORMATIVE EVALUATION POLICIES**

The nature of an internal evaluation unit means the SEU has a professional, ethical, and moral imperative to not only manage the process of making judgments about if OCB is doing things right, but also manage the process of asking if OCB is doing the right things. The first type of judgment is related to effectiveness principles, the latter is related to moral principles. Both can and should be evaluated. For instance, evaluation can be an extractive, invasive, and a harmful process to individuals and communities who need life-saving medical humanitarian interventions. When this happens, it is a moral wrong. Thankfully there was no evidence of harm or wrongdoing in past evaluation cases identified through the course of this meta-evaluation. And still, there is an opportunity for the SEU to be a leader within the MSF movement to model what it looks like to wrestle with and address historical and systematic injustices bound up in the humanitarian sector by addressing how these injustices may manifest through evaluation practice. With that, the SEU should give serious consideration to a new cluster of evaluation policies that extend cosmopolitan notions of ethics to include a vision of transformative evaluation practice that is culturally responsive, culturally specific, and equitable for those involved in and affected by the evaluation function.

Recommendations 4-5 (of 5)

>-<

ANNEXES

ANNEX I: TERMS OF REFERENCE

Médecins Sans Frontières (MSF)/Doctors Without Borders is an international medical humanitarian organization determined to bring quality medical care to people in crises around the world, when and where they need regardless of religion, ethnical background, or political view. Our fundamental principles are neutrality, impartiality, independence, medical ethics, bearing witness and accountability.

The Stockholm Evaluation Unit (SEU), based in Sweden, is one of three MSF units tasked to manage and guide evaluations of MSF's operational projects. For more information see: evaluation.msf.org.

Project:	Meta Evaluation
Start/end date:	May 2022 – October 2022
How to apply:	Interested candidates are invited to submit: 1) A proposal describing how the evaluation will be conducted (including a budget in a separate file) 2) CV (s) 3) A written example of an evaluation already carried out
Deadline to apply:	24th April, 2022
Application to be submitted to:	evaluations@stockholm.msf.org
Specific considerations:	The proposal must contain a suggestion of the most appropriate and available criteria that will be used as a basis for this evaluation. This will then be finalized as a part of the inception phase.

BACKGROUND

The commitment to evaluation at MSF comes primarily from the La Mancha Agreement (2006)[73] which states not only that MSF aspires to ensure quality, relevance, and extent of operations, and to commits to the impact and effectiveness of its work so that good work can be multiplied, and ineffective practice abandoned. MSF Operational Centre Brussels (OCB) elaborates on this commitment in its 2020-23 Strategic Orientations stating that it wants to develop: a culture of evaluation to give the field teams the opportunity to learn from [their] practices and to constantly improve the quality and pertinence of operational/medical interventions.

MSF does not accept institutional funding from most bilateral donors, removing what is often an impetus for evaluation at other non-governmental organizations. Learning is most often cited as the

[73] The La Mancha Agreement was adopted in Athens, Greece in 2006 following a process of discussion and debate to address internal challenges. https://msf.org/sites/msf.org/files/La%20Mancha%20Agreement%20EN.pdf.

predominant intention behind wanting to evaluate. This can be related directly to the individual project, future programing in the country or region, to inform advocacy (vis-à-vis for example a country's ministry of health) as well institutional learning.

For OCB, evaluation is about assessing the design, strategy, implementation, and results of medical and humanitarian interventions, measured against established MSF or international standards (SEU Steering Committee Framework, 2019). A dedicated unit, the Stockholm Evaluation Unit (SEU), manages primarily external evaluations, but does on occasion conduct internal evaluations as well. They cover a range of medical operational topics (i.e., migration, non-communicable disease, HIV/AIDS), and in some cases topics related more to organizational sets-up and strategies.

There is currently no formal adopted framework of quality in the evaluations managed by the SEU on behalf of OCB although the work of the unit is influenced by several frameworks including the Joint Committee on Standards for Educational Evaluation (JCSEE) Program Evaluation Standards, ALNAP Proforma, as well as various evaluator competency frameworks, including those from the American Evaluation Association and the United Nations Evaluation Group (UNEG). It is likely that ideas on what constitutes quality or value for different stakeholders in evaluation within the context of MSF and OCB differ across the organization. It will be necessary to establish a framework of accepted criteria as part of the evaluation process.

This meta evaluation will seek to assess completed evaluations carried out between 2017-2022[74], that the SEU has managed at the request of or directly and significantly involving OCB. Other entities at OCB do complete their own internal, analytical exercises (i.e., retrospectives) but these would be out of scope for this evaluation.

PURPOSE AND INTENDED USE

The purpose of this meta evaluation is to assess the quality and value of OCB evaluations. The intention is not to evaluate the SEU's performance but rather the evaluations (individually and collectively) that have been finalized at OCB's request, and the unit has managed. This should not be a technocratic exercise, based on checklists that review whether specific elements (i.e., inception report) have been included, but rather an analytical exercise that assesses the value of the completed evaluations to OCB, ranging from individual projects to the organization as such.

This meta evaluation should help to build a coherent understanding of what constitutes value and quality of evaluations to OCB. Results should explore factors influencing the value and quality of evaluation, and how these can be increased within the organizational context. Understanding evaluations' significance can contribute to shed light on their worth. The primary recipients of the meta evaluation are the SEU and the SEU Steering Committee; the secondary recipients are the OCB Board, OCB association and staff.

As stated, the SEU does today not manage a formalized quality framework to define what is quality and value in evaluations at OCB. That said, it is guided by three overarching areas (methods, use and values) that can serve as subheadings for standards. The table below provides examples of criteria but is not exhaustive.

	Method	Use	Value

[74] This constitutes roughly 28 evaluations and other evaluative exercises.

Examples of criteria	• Credible • Accuracy • Feasibility • Professional integrity	• Utility	• Principles • Ethics • Human rights

The proposal must make suggestions of the most appropriate criteria to be used, which will then be elaborated upon and finalized as a part of the inception phase.

EXPECTED DELIVERABLES

1. Inception Report

The inception report ought to include a detailed evaluation proposal including the methodology and evaluation protocol. The IR must elaborate on the evaluand and evaluation questions and include the proposal of criteria to be used to assess quality.

2. Draft Evaluation Report

The draft ER ought to answer to the evaluation questions and will include analysis, findings, and conclusions – and if necessary – lessons learned and recommendations.

3. Working Session

As part of the report writing process, a working session will be held with the commissioner, consultation group members and SEU evaluation manager. The evaluator will present the preliminary findings, collect feedback and facilitate a discussion on recommendations (either to co-create recommendations or, if already developed, their feasibility).

4. Final Evaluation Report

The final report will have addressed feedback received during the working session and written input from the feedback loop.

5. Presentation of the Final Evaluation Report

A presentation of the final report to a general OCB audience in the form of a webinar.

The key deliverables (inception report, draft/final report) will be processed through a feedback loop, collecting input from the consultation group (see below, Practical Implementation of the Evaluation). They are then endorsed by the evaluation's commissioner.

TOOLS AND METHODOLOGY PROPOSED

In addition to the initial evaluation proposal submitted as a part of the application, a detailed evaluation protocol should be prepared by the evaluators during the inception phase. It will include a detailed explanation of proposed methods and its justification based on validated theories. It will be reviewed and validated as a part of the inception phase in coordination with the SEU.

RECOMMENDED DOCUMENTATION

- Evaluations and other evaluative exercises managed by the SEU for OCB 2017-2022
- Existing SEU plans, guidelines, and policies

- SEU Steering Committee framework (2019)
- OCB Strategic Orientation 2020-2023, OCB Strategic Prospects 2020-2023

PRACTICAL IMPLEMENTATION OF THE EVALUATION

Number of evaluator(s)	Flexible
Timing of the evaluation	start May 2022 - finish September

The SEU and its steering committee will establish a consultation group (CG) to accompany this evaluation. The CG is led by a commissioner. They have contributed to finalizing this ToR.

PROFILE/REQUIREMENTS FOR EVALUATOR(S)

The evaluation requires an individual or team of individuals who can demonstrate competencies in the following areas.

1. Relevant evaluation competencies, preferably with experience in implementing a meta-evaluation like the one being proposed
   a. Professional focus - acts ethically, reflectively, enhances and advances professional practice of evaluation.
   b. Technical focus - applies appropriate evaluation methodology.
   c. Situational focus - considers and analyses evaluation context successfully.
   d. Management focus - conducts and manages evaluation projects skillfully.
   e. Communication focus - interacts and communicates successfully with stakeholders.
2. Technical competencies
   a. Humanitarian program management, including humanitarian program monitoring and evaluation and/or knowledge management and learning.
   b. Fluency in English, spoken and written. French is a benefit.

APPLICATION PROCESS

The application should consist of a technical proposal in English, a budget proposal, CV, and a previous work sample. The proposal should include a reflection on how adherence to ethical standards for evaluations will be considered throughout the evaluation. In addition, the evaluator/s should consider and address the sensitivity of the topic at hand in the methodology as well as be reflected in the team set-up. Offers should include a separate quotation for the complete services, stated in euros. The budget should present consultancy fee according to the number of expected working days over the entire period, both in totality and as a daily fee. Travel costs, if any, do not need to be included as the SEU will arrange and cover these. Do note that MSF does *not* pay any per diem. Applications will be evaluated based on whether the submitted proposal captures an understanding of the main deliverables as per this ToR, a methodology relevant to achieving the results foreseen, and the overall capacity of the evaluator(s) to carry out the work (i.e., inclusion of proposed evaluators' CVs, reference to previous work, certification et cetera).

Interested teams or individuals should apply to **evaluations.sweden@stockholm.msf.org** referencing **[META]** no later than **Sunday April 24, 23:59 CET.** We would appreciate the necessary documents being submitted as separate attachments (proposal, budget, CV, work sample and such). Please include your contact details in your CV. Please indicate in your email application on which platform you saw this vacancy.

ANNEX II: DETAILED METHODS NOTE

The following methods note uses the Checklist for Evaluation-Specific (reporting) Standards (CHESS) to present the "minimum, evaluation-specific elements that must be reported to make judgments about the quality of the evaluation" and theoretically replicate procedures, as needed.

Domain	Category	No.	Category Values
People/ Personnel	Evaluator(s)	1	
	Affiliation	1a	• Zach Tilton: Pointed Arrows Consulting (Contracted Firm); Interdisciplinary PhD in Evaluation Program, Western Michigan University • Tian Ford: Pointed Arrows Consulting (Contracted Firm); Joan B. Kroc School of Peace Studies, University of San Diego • Dr. Michael Harnar: Pointed Arrows Consulting (Contracted Firm); Interdisciplinary PhD in Evaluation Program and The Evaluation Center, Western Michigan University
	Disciplinary Training	1b	• Zach Tilton: BS: Peacebuilding; MA: Peace Studies; PhD Evaluation (in process) • Tian Ford: BA: Peacebuilding; MS: Conflict Resolution and Management (in process) • Dr. Michael Harnar: BGS (Bachelor of General Studies); MA: Psychology; PhD Psychology, Emphasis in Evaluation and Applied Research Methods
	Role	1c	• External meta-evaluation team
	Gender	1d & e	• Zach Tilton: cis-gendered male • Tian Ford: BA: cis-gendered male • Dr. Michael Harnar: cis-gendered male
	Ethnicity	1f	• Zach Tilton: white, American • Tian Ford: BA: white, American • Dr. Michael Harnar: white, American

	Years experience in evaluation	1g	• Zach Tilton: 10 years • Tian Ford: 1 year • Dr. Michael Harnar: 20 years
	Languages Used	1h	● English, at least one instance of machine-assisted translation of French document
	Epistemological Orientation	1i	• Zach Tilton: pragmatist, constructivist • Tian Ford: pragmatist, constructivist • Dr. Michael Harnar: pragmatist, constructivist
	Funder(s)	2	• Médecins *Sans Frontières,* Operational Centre Brussels,
	Client(s)	3	• Médecins *Sans Frontières,* Operational Centre Brussels, • Commissioners: Marc Biot Operations Director & Catherine Van Overloop, Medical Director
	Audience(s)	4	• Funders/investors, directors/managers, service providers, patients, program designers, consultants, scholars, policy makers, evaluation practitioners, general public
	Relevant Stakeholders	5	• Stockholm Evaluation Unit, meta-evaluation consultation group, Stockholm Evaluation Unit Steering Committee, Operations and Medical Department Directors and Managers, prior evaluation consultants, managers, commissioners, and project contacts; cell, country, and project-level staff; patients and communities
	Primary Stakeholders	6	• Stockholm Evaluation Head of Unit, evaluation managers, meta-evaluation consultation group, Stockholm Evaluation Unit Steering Committee, Operations and Medical Department Directors and Managers

Domain	Category	No.	Category Values
Evaluation	Evaluation Type	7	• Summative retrospective portfolio meta-evaluation

Context and Characteristics	Evaluand Type	8	• Portfolio of evaluation cases comprising performances and products; centralized internal evaluation unit and evaluation system comprising performances, processes, and personnel
	Substantive Area	9	• Medical humanitarian evaluation
	Funding Type	10	• Competitive, RFP/RFA/RFC
	Date(s) Evaluation Commissioned	11	• May 2022
	Date(s) Evaluation Conducted	12a	• May 2022
		12b	• Dec 2022
	Geopolitical Scope	13	• Global; Remote
		14	• Multi-nation, multi-site
	Scale	15	• Number and size of site(s)
	Political context	16	• Political hostility was low; internal evaluation unit funding was not under question; organization is generally favorable to evaluation rhetorically and mostly in practice with some skeptics or uninitiated; emergent evaluation culture at the operational centre; there are reported philosophical differences between other internal evaluation units at other centres across the MSF movement. However, given the scope, nature, and potential implications of the meta-evaluation, the stakes were moderate to high.

Domain	Category	No.	Category Values

Investigation design and methods	Evaluation Purpose	17	• First and foremost to determine merit and worth of evaluation portfolio and system; secondary purpose for use improving
	Evaluation Approach	18	• Stufflebeam's meta-evaluation checklist approach
	Procedure(s) For Identifying Stakeholders	19	• Evaluation participants were identified beforehand by the evaluation manager through the formation of a consultation group and the selection of inception stage key informants and consultation groups. • Survey respondents were identified by number and nature of roles per evaluation case and then identified with the actual number of individuals who fit those roles. • Intended users were identified and discussed with evaluation manager using a utilization focused inspired workbook for such an activity.
	Procedure(s) For Prioritizing Stakeholders	20	• Intended user groups and users were identified before evaluation team involvement, but refined with the use of an intended user workbook presented to the evaluation manager/head of unit.
	Procedures For Engaging Stakeholders	21	• Key informant interviews, focus group discussions, emails, weekly meetings with evaluation manager, online survey, pause and reflect workshop, sensemaking workshops, webinar presentations
	Valuing process	22a	Sources of criteria: • OCB staff professional values • SEU Evaluation Policies • Generally Accepted Evaluation Quality Frameworks
		22b	Procedure(s) for establishing criteria • Suggestion of quality domains in ToR (Values, Methods, Use) • Proposal from meta-evaluation team of quality frameworks • Refinement of frameworks after inception stage organizational values inquiry
		22c	Procedure(s) for prioritizing criteria

			• Priority ranking of the quality frameworks and five Program Evaluation Standards criteria were conducted with the consultation group. Process of inviting the consultation group to consider relative rank of frameworks and criteria did not factor into the final report, though differential weighting was applied at one point, the weighting did not create significant variation in scores and given low representation in ranking exercise participants, efforts for differential criteria weighting were abandoned.
		22d	Procedure(s) for establishing standards: • Generic standards were used for all four quality frameworks as following: ○ Criteria: poor, fair, good, very good, excellent; ○ Sub-criteria: poor, fair, good, very good, excellent; ○ Indicators (PrgES and ALNAP): met, met* (not applicable), not met, not met* (no evidence) • The PrgES and ALNAP frameworks had standards differentiated by cut-scores established by the PrgES checklist procedures; UNEG and EMQF used a generic rubric developed by the meta-evaluation team. • These were presented to the consultation group, but the meta-evaluation team owned the whole process for establishing standards • Applying the EMQF and UNEG frameworks followed these steps: gather multiple pieces of information from multiple sources to describe a dimension that is expressed in the framework and compare the developed description of practice with the ideal expressed in the framework. The more the actual performance matches an ideal performance, the higher the performance rating. The less the actual performance matches an ideal performance, the lower the performance rating. • This process of description, comparison, and rating was repeated for each framework sub-criterion (the smallest unit of analysis for these frameworks were sub-criteria). Sub-criteria ratings were then assigned a numerical value (poor=1; fair=2; good=3; very good=4; excellent=5) and averaged within a dimension to determine the overall dimension rating, rounding to the nearest whole number value and rounding up at half points. Criteria domain rating values were then averaged for an overall framework rating. • There are three levels of data in the PrgES and ALNAP: the upper two are derived from the lowest indicator, summing up through sub-criteria to create a rating on a criteria. Cut scores for each level: Poor (0-16%), Fair (17-41%), Good (42-66%), Very Good (67-91%), and Excellent (92-100%). Stufflebeam (2016) says "There is no magic formula for setting cut scores that mark boundaries of the five rating categories. The set [incorporated into the checklist] reflects the checklist author's judgments, based on several previous metaevaluations but are only one of many options. At first glance, the…ranges may seem lenient, but they aren't….Users of the checklist may apply this checklist's determined rating ranges or, instead, thoughtfully, systematically, and transparently set an alternative set of cut points."

	Sample	23	• 31 evaluation cases from 2017-2021; • Consultation group; steering committee; and select focal headquarter points for key informant interviews and focus groups • A survey response rate of 42% (n= 57/133) was decent for an external survey and likely low for an internal survey. Managers comprised the largest respondent group with 42% of total responses (n=24/57), then evaluators with 32% (n= 18/57), followed by commissioners with 16% (n= 9/57) and project contacts with 11% (n= 6/57).
	Sampling Procedure(s)	24	• Evaluation case sampling procedure was purposive sample of SEU managed evaluation cases for OCB; • Online survey sampling procedure was a purposive stratified sample by evaluation case role (evaluator, manager, commissioner, project contact)
	Procedure(s) For Establishing Questions	25	• Illustrative meta-evaluation questions were posed by the meta-evaluation team in the proposal • Meta-evaluation questions and sub-questions were finalized through consultation in the inception phase and agreed in the inception report. There was little deviation between proposed and final questions.
	Procedure(s) For Prioritizing Questions	26	• Interpretation of ToR; consultation with primary intended users; agreement among evaluation team members
	General Methodological Orientation	27	• Mixed method
	Research Design	28	• Mixed method sequential synthesis design. The specific meta-evaluation design is an external practical summative meta-evaluation subsequent to evaluation performances using multiple data sets where original evaluation data are not manipulated.
	Data Collection Instruments	29a	• Key informant interview and focus group discussion protocols; online survey protocols; evaluation checklist extraction form

		29b	Proprietary instruments: • Program Evaluation Meta-evaluation Checklist; ALNAP Proforma; UNEG Norms & Standards

Domain	Category	No.	Category Values
Evaluative Argument and Conclusions Domain	Results	30	• Presented in order by meta-evaluation question; where applicable, scores, ratings, and ranks provided for evaluation criteria and cases; definitions, descriptions, and judgments provided for most or select criteria
	Synthesis procedures	31	• Interview and focus group data were transcribed and coded by scheme pertaining to values, criteria, evidence quality, and intended use; • Numeric weight and sum methodology, specifically the procedure outlined in the program evaluation meta-evaluation checklist[75], replaced the grading synthesis prescribed in the ALNAP proforma; and qualitative evaluation rubrics were applied for UNEG and EMQF frameworks with crude numeric weight and sum methodology for numerically coding and averaging ratings for sub-criterion synthesis. • Combined, the PrgES and the ALNAP Proforma provided 223 qualitative indicators of evaluation quality judged "met" or "not met." These binary judgements rolled up into qualitative ratings and quantitative scores for 45 sub-criteria of evaluation quality for each case. These 45 sub-criteria rolled up into qualitative ratings and quantitative scores for 10 overarching criteria and quality domains for each case. Average scores for these 10 criteria and domains resulted in overall scores, ratings, and ranks for each evaluation case. In addition to computing scores, ratings, and ranks for each evaluation case, scores, ratings, and ranks were derived for each of the 5 overarching criteria in the PrgES and 5 overarching quality domains in the ALNAP Proforma across all cases for portfolio-level scores, ratings, and ranks for these dimensions of quality. • Combining the 223 indicators from the PrgES and ALNAP frameworks across 31 evaluation cases translated into 6,913 qualitative ratings that corresponded to 45 sub-criteria and 10 overarching criteria. • General evidence sources for indicator judgments across PrgES and ALNAP frameworks are: evaluation artifacts, online survey data, key informant interview and focus group discussion transcripts, and participant observation notes.

[75] We used a 2016 private proprietary version of the checklist, but to see the exact formulae for synthesis, see this public version adapted by USAID.

- The ALNAP Proforma suggests grading evaluations with dual raters, but does not include a rubric to prescribe how to transparently and systematically assign grades, let alone reliable grades between dual raters. It also does not explain how to resolve grading conflicts between two reviewers. We addressed this with the reliability analyses reported below.

- Additional analytical procedures were conducted including:

  - PrgES record to modified EMQF categories, where we re-coded each of the 223 PrgES and ALNAP indicators by the three main quality domains of *Value*, *Use*, and *Method* as well as a fourth domain of *Transformation*. This re-coding created the ability to view 6,913 data points across 31 cases by scores, ratings, and ranks for the EMQF quality domains.

  - Cost-utility analysis where standardized z-scores for Utility measures were compared with standardized z-scores for evaluation case budgets.
  - Use and influence analysis consisting of categorization and classification of use and use outcome reports, as indicated in the use and use outcome section.
  - Historical analysis, where evaluation case quality scores were correlated with years to demonstrate an association.
  - Six-step analysis: Applying the SEU's Six-step process model as an analytical tool involved a process where core calculations did not reweight across the step domains like the standard PrgES and ALNAP formulas. This means that steps with fewer number of checkpoints or indicators, actually have a higher weight than those with more indicators. This is because scores for each domain are derived from percentages of met/not met for that domain. The total evaluation step score is then an average of these scores, which means domains with fewer checkpoints have more sensitivity or potential for variation in percentages due to smaller denominators for the total number of possible checkpoints per domain.
  - Maximum deviation analysis, where open ended survey responses about factors of quality for highest and lowest performing evaluations were coded thematically.
  - Limitation analysis where most frequently reported limitations from limitation section of all 31 evaluations were coded and reported.
  - Role and responsibility analysis where open-ended responses where evaluation participants by role were asked to describe the responsibility someone in their role held for evaluation quality. Text segments were coded thematically.

	Comparison procedures	32	• Absolute comparisons were made between generally accepted quality evaluation standards and actual evaluation performances; relative comparisons were made between cases within the portfolio; no external relative comparisons were made with comparable evaluation portfolios of other operational centres, other medical humanitarian agencies, or other aid agencies.
	Interpretation process	33	• The initial interpretation and sensemaking process was conducted internally among meta-evaluation team members after initial data extraction and collection processes were completed. • Findings were initially presented by a report draft to the head of unit; initial feedback from the head of unit was integrated before presenting second draft to SEU managers consultation group and steering committee. All evaluation managers provided detailed feedback, with some consultation group and steering committee members providing general feedback; responses to detailed and general feedback were given; final draft was delivered after additional data analysis on use and use outcomes was conducted. • A sensemaking workshop with the primary intended user group, the SEU, was conducted with the online tool Miro board to make meaning for most significant findings. • Oral presentations were given during an online video call to the steering committee and then again to the OCB staff during a lunch and learn webinar.
	Limitations	34	• This meta-evaluation had a few limitations that were known and unknown at the start of the work. Known issues were tight timeframes for reaching framework consensus; large differences in time-zones between evaluation team members and clients (GMT -10 and GMT +1 at the extremes); lack of French speaker on evaluation team; positionality blind spot in homogenous lived and privileged experiences of evaluators; inter-rater reliability issues without evaluation case dual-rating; and assumed philosophical reservations among some users to the proposed checklist and numerical weight and sum methodologies as potentially too technocratic. Of all of these, time zone differences and timeframes had the most effect in the first four evaluation stages. Timeframe issues were more operational, or more apparent, in data collection and analysis. The scoped and final meta-evaluation framework was known to be ambitious and even so, level of effort forecasts for data extraction, analysis, and reporting were severely underestimated. This translated to compressed and fast-tracked interpretation and reporting procedures resulting in foregone sophisticated data visualizations in the final report. Instrumentation challenges occurred when pivoting from email interviews to online surveys, which translated into delays and unsatisfied survey respondents. Self-selection, courtesy, self-serving, recall, and social acceptability biases are all possibilities with the self-report data from our purposefully sampled online survey about past evaluation performances. A survey response rate of 42% (n= 57/133) was decent for an external survey and likely low for an internal survey. Managers comprised the

			largest respondent group with 42% of total responses (n=24/57), then evaluators with 32% (n= 18/57), followed by commissioners with 16% (n= 9/57) and project contacts with 11% (n= 6/57). Relative to inputs, processes, and outputs, outcome-level data was lacking due to limited evaluation follow-up and use documentation. Finally, although two subject-matter experts (SME) for medical evaluation and humanitarian evaluation were successfully recruited, engagement with these SMEs was limited and of low influence in the design and execution of the meta-evaluation and ultimately no response and integration of their feedback toward the end of the process when provided with report drafts and technical questions. The meta-evaluation team believes these limitations were sufficiently addressed during the conduct and or accounted for in this final report and do not pose undue threats to the validity of meta-evaluation conclusions.
	Statement of conclusions	35	• Each meta-evaluation question has findings that include evaluative conclusions about dimensions and sub-dimensions of quality. • Overall conclusions for key meta-evaluation questions and other findings are presented at the end of the report; operational recommendations were provided for each finding in the main report body; strategic recommendations were provided after key conclusions.

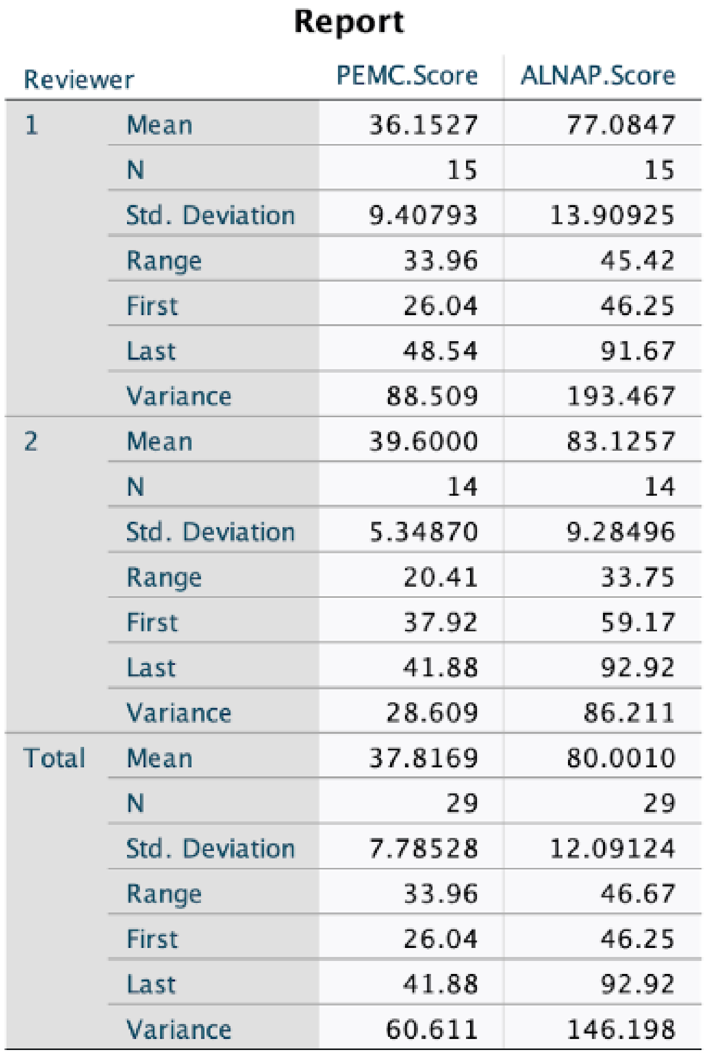
DETAILED METHODS SUB-NOTE ABOUT RELIABILITY ANALYSES

Inter-rater alignment: Given the breadth of synthesized meta-evaluative frameworks and the number of evaluation artifacts, independent dual-rating of each evaluation was untenable. This has implications for reliability. To address this, reviewers developed a plan for initial calibration of evaluation rating, and intermittent checks on issues encountered, interpretation, and quality. Two raters split the coding and review of the catalog of documents. Concerns over aligning ratings were addressed by having the coders work through a single evaluation's set of documents and complete both the PrgES and ALNAP instruments. This first pass provided a moderately acceptable agreement (PEMC = 55%; ALNAP = 65%). They discussed at length each unaligned coding before they reviewed and coded a second evaluation. The agreement increased to a percentage that was deemed adequate for this project (PEMC = 83%; ALNAP = 94%).

For transparency sake, coding choice assumptions were made explicit throughout the coding process, using a comments column in the coding sheet. This proved useful in the reviews described next.

Qualitative review: The coding spreadsheet was sorted by the reviewer and each indicator coding was reviewed looking for unusual coding patterns or where a code was almost entirely unmet/met. This qualitative analysis fed into the systematic coding choices mentioned elsewhere. For example, one of the reviewers had been assigned 4–5 evaluations that were assumed to not necessarily be humanitarian and that assumption had made itself known through how a few codes were applied by that reviewer differently from the other. Once this had been uncovered, an assumption that specific indicators did not apply was interrogated and the coding was revised.

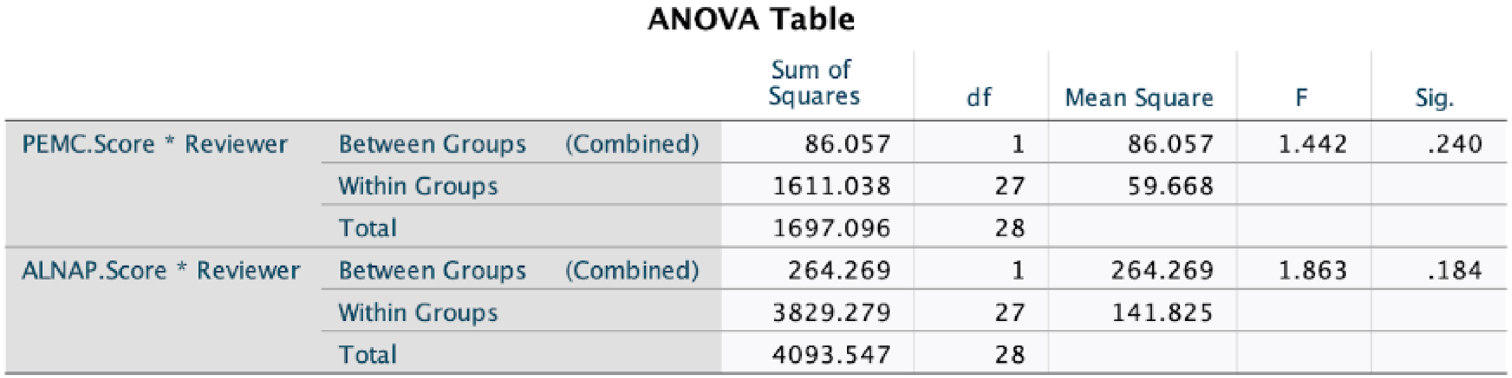
Quantitative analysis: The average assessment scores for each evaluation were compared across reviewers and scores for each instrument were tested for correlations. There is a significant positive correlation between the PrgES and ALNAP ratings (r = .460, p = .009). Differences between reviewers were not significant for either the PrgES (f = 1.444, p = 2.40) or the ALNAP (f = 1.863, p = .184).



Report

Reviewer		PEMC.Score	ALNAP.Score
1	Mean	36.1527	77.0847
	N	15	15
	Std. Deviation	9.40793	13.90925
	Range	33.96	45.42
	First	26.04	46.25
	Last	48.54	91.67
	Variance	88.509	193.467
2	Mean	39.6000	83.1257
	N	14	14
	Std. Deviation	5.34870	9.28496
	Range	20.41	33.75
	First	37.92	59.17
	Last	41.88	92.92
	Variance	28.609	86.211
Total	Mean	37.8169	80.0010
	N	29	29
	Std. Deviation	7.78528	12.09124
	Range	33.96	46.67
	First	26.04	46.25
	Last	41.88	92.92
	Variance	60.611	146.198

Instrument coverage: Reliability of the instrument applications and how they cover different topics was explored by graphing the scores on a standard continuum and doing a qualitative review. We constructed z-scores for each PrgES and ALNAP score. The z-score function places each report on a continuum that is relative to the mean of the scores on that index (e.g., PrgES mean) and the standard deviation around that mean. These z-scores were then sorted by the PrgES and graphed together. The mid-point of the x axis is the mean of PrgES scores. The graph gives a glimpse of the relative relationship between the PrgES scores and the ALNAP scores. Easily highlighted are those evaluations that scored, what looks to be a dramatic difference (e.g., opposite) on the two indexes and how that relates to differences in coverage of the PrgES and the ALNAP.



ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
PEMC.Score * Reviewer	Between Groups	(Combined)	86.057	1	86.057	1.442	.240
	Within Groups		1611.038	27	59.668		
	Total		1697.096	28			
ALNAP.Score * Reviewer	Between Groups	(Combined)	264.269	1	264.269	1.863	.184
	Within Groups		3829.279	27	141.825		
	Total		4093.547	28			

DETAILED METHODS SUB-NOTE ABOUT META-EVALUATION TEAM COMPOSITION

The team of meta-evaluators was composed of three evaluation and meta-evaluation practitioners and researchers of meta-evaluation. From emerging, early to mid-career, and well-established, the practitioners on the team have over 30 years of evaluation and meta-evaluation experience. As Senior Meta-evaluation Advisor, Dr. Michael Harnar is an evaluation consultant and the Interim Director of the Interdisciplinary PhD in Evaluation program (IDPE) at Western Michigan University where his research agenda focuses on evaluation quality (e.g. Harnar, Hillman, Endres, and Snow, 2020), meta-evaluation theory and practice and evaluation use. As a Lead Meta-evaluator, Zach Tilton is a doctoral candidate in the IDPE and an evaluation consultant where his research and consulting practice focuses on peacebuilding evaluation, meta-evaluation, and technology-enabled evaluation. As a Junior Meta-evaluator, Tian Ford is an evaluation consultant who currently works with community-based organizations on evaluability assessments and evaluation capacity development. Dr. Harnar and Mr. Tilton have written about the nature of and key issues within meta-evaluation (Wingate, Tilton, and Harnar, 2023, in press) and are currently conducting transdisciplinary research on meta-evaluation practice. They are also researchers on evaluation, where they study evaluation use, at The Evaluation Center at Western Michigan University, which houses the Joint Committee on Standards for Educational Evaluation.

With minor variation between practitioners on the team, all three team members are ontological pragmatists (Mertens and Wilson, 2019; Patton and Campbell-Patton, 2022) and historical realists (Lincoln and Guba, 2005). That is to say we are less concerned with resolving questions about metaphysics and more interested in the difference metaphysical assumptions make in the lives of evaluation users, constituents, and in their organizations. We believe there may be a single reality independent from knowers, but that individual knowers have their own unique interpretation of reality and that those interpretations form historical realities that are shaped by social, political, cultural, economic, ethnicity and gender values; crystallized over time and have real consequences. Further, we believe the main function of evaluation should be on valuing (Scriven, 1991; Schwandt, 2015) in service of public good and social transformation (Mertens, 2008). While there is less variation in gender, race, and language on our team, we believe in interrogating how our lived experiences manifest in potential biases in the meta-evaluative process, ensuring the meta-evaluation is both equitable and culturally responsive, and in determining the extent to which the evaluations under review manifest those ideals as well.

All team members' professional values are exemplified by the guiding principles authored by the American Evaluation Association (2018): Systematic inquiry: conduct data-based inquiries that are thorough, methodical, and contextually relevant; Competence: provide skilled professional services to the project stakeholders; Integrity/honesty: behave with honesty and transparency to ensure the integrity of the evaluation; Respect for people: honor the dignity, well-being, and self-worth of individuals and acknowledge the influence of culture within and across groups; Common Good and Equity: strive to contribute to the common good and advancement of an equitable and just society.

ANNEX III: PROGRAM EVALUATION STANDARDS CRITERIA SUMMARY

Evaluation Code	Utility Score	Feasibility Score	Evaluation Accountability Score	Propriety Score	Accuracy Score	Total Score	Total Rating
PORTFOLIO	**68%**	**66%**	**63%**	**53%**	**51%**	**60%**	**Good**
MUMPO	91%	88%	67%	72%	75%	**78%**	**Very Good**
BOLIM	84%	88%	67%	78%	75%	**78%**	**Very Good**
MBADO	84%	94%	67%	81%	59%	**77%**	**Very Good**
NCDKE	84%	88%	67%	53%	75%	**73%**	**Very Good**
BILIC	88%	94%	67%	66%	63%	**75%**	**Very Good**
GUCCE	91%	94%	67%	63%	53%	**73%**	**Very Good**
REACH	91%	88%	67%	72%	53%	**74%**	**Very Good**
ESHIV	81%	88%	67%	69%	59%	**73%**	**Very Good**
EBOLA	88%	81%	67%	72%	63%	**74%**	**Very Good**
EVAL21	84%	94%	67%	75%	50%	**74%**	**Very Good**
ARCHE	84%	81%	58%	72%	59%	**71%**	**Very Good**
OCHMU	91%	81%	58%	50%	53%	**67%**	**Good**
EPOOL	72%	69%	58%	59%	69%	**65%**	**Good**
DGDFM	75%	81%	58%	75%	47%	**67%**	**Good**
HIVKIN	69%	81%	67%	72%	38%	**65%**	**Good**
COMME	75%	81%	67%	47%	50%	**64%**	**Good**
IDAII	72%	69%	58%	69%	50%	**64%**	**Good**
SUPCH	75%	63%	67%	59%	63%	**65%**	**Good**
HREVA	69%	81%	67%	38%	53%	**61%**	**Good**
VOTTR	84%	63%	67%	41%	47%	**60%**	**Good**
DIGHP	72%	75%	67%	38%	50%	**60%**	**Good**
USCOV	50%	31%	67%	59%	50%	**51%**	**Good**
FRCOH	50%	50%	67%	47%	47%	**52%**	**Good**
VTCAR	66%	38%	50%	44%	50%	**49%**	**Good**
MASTE	53%	56%	67%	31%	44%	**50%**	**Good**
OCBFE	41%	25%	58%	25%	53%	**40%**	**Good**
BUDGE	44%	31%	58%	28%	38%	**40%**	**Good**
EMRKS	25%	31%	67%	22%	13%	**31%**	**Fair**
MVGCE	25%	19%	58%	25%	28%	**31%**	**Fair**
OCBPR	28%	25%	50%	22%	25%	**30%**	**Fair**
MAURT	13%	19%	50%	13%	16%	**22%**	**Fair**

ANNEX IV: ALNAP PROFORMA QUALITY DIMENSIONS SUMMARY

Eval Code	ToR Score	Methods Score	Contextual Analysis Score	Intervention Assessment Score	Report Assess-ment Score	Overall Score	Overall Rating
PORTFOLIO	**82%**	**72%**	**80%**	**78%**	**88%**	**77%**	**Very Good**
EBOLA	88%	83%	100%	94%	100%	93%	Excellent
MUMPO	88%	83%	100%	88%	100%	91%	Very Good
VTCAR	75%	67%	100%	100%	100%	88%	Very Good
MBADO	75%	83%	100%	88%	100%	87%	Very Good
REACH	100%	83%	100%	63%	100%	87%	Very Good
EPOOL	63%	83%	100%	94%	100%	87%	Very Good
MASTE	75%	83%	100%	88%	100%	86%	Very Good
BILIC	88%	83%	100%	75%	88%	86%	Very Good
IDAII	75%	83%	100%	75%	100%	86%	Very Good
BOLIM	100%	75%	100%	81%	75%	85%	Very Good
OCHMU	100%	83%	100%	81%	100%	84%	Very Good
EVAL21	88%	50%	100%	88%	100%	83%	Very Good
DGDFM	88%	83%	75%	69%	100%	83%	Very Good
FRCOH	88%	75%	100%	88%	88%	81%	Very Good
BUDGE	88%	83%	100%	69%	100%	80%	Very Good
USCOV	63%	67%	100%	75%	100%	79%	Very Good
GUCCE	75%	92%	50%	88%	100%	79%	Very Good
OCBFE	88%	67%	100%	81%	100%	78%	Very Good
HREVA	88%	67%	100%	81%	75%	78%	Very Good
SUPCH	88%	67%	100%	81%	100%	78%	Very Good
ESHIV	88%	67%	75%	63%	100%	75%	Very Good
ARCHE	88%	75%	25%	81%	88%	71%	Very Good
DIGHP	88%	58%	100%	44%	75%	71%	Very Good
OCBPR	75%	33%	100%	100%	88%	70%	Very Good
COMME	88%	83%	0%	88%	100%	69%	Very Good
MVGCE	63%	58%	100%	63%	63%	68%	Very Good
NCDKE	88%	83%	25%	81%	75%	66%	Good
HIVKIN	63%	83%	50%	69%	63%	65%	Good
VOTTR	75%	83%	0%	63%	75%	57%	Good
MAURT	75%	17%	75%	56%	38%	52%	Good
EMRKS	75%	50%	0%	69%	38%	45%	Good

ANNEX V: COMBINED OVERALL ALNAP & PRGES SCORES, RATINGS, AND RANKS FOR EVALUATION CASES

The following table reports and averages the total scores from the PrgES and ALNAP frameworks for updated combined scores, ratings, and ranks for evaluation cases and the evaluation portfolio. The idea is that the PrgES and ALNAP are both robust, but partial frameworks, or investigate mostly similar, but also slightly different dimensions of quality. Given this, it may be these combined scores, ratings, and ranks are the most accurate quality judgements for cases. Despite the updated portfolio quality score results in a "Very Good" rating, the main evaluation conclusion for MEQ2 about evaluation quality still remains that past evaluations are "Good" to "Very Good" given the UNEG and EMQF frameworks also varied between those two ratings, and have a less defensible basis for averaging those overall ratings given those frameworks and their application in this study did not have the same level of transparency and replicability as the PrgES and ALNAP frameworks. For the most detailed report of evaluation case quality readers can refer to METAE Dashboard, which contains detailed PrgES and ALNAP scorecards.

Eval Code	ALNAP Score	PrgES Score	Combined Score	Combined Rating	Combined Rank
Portfolio	**77%**	**60%**	**69%**	**Very Good**	**NA**
MUMPO	91%	77%	84%	Very Good	1
EBOLA	93%	71%	82%	Very Good	2
MBADO	87%	75%	81%	Very Good	3
BOLIM	85%	76%	81%	Very Good	4
REACH	87%	72%	80%	Very Good	5
BILIC	86%	73%	79%	Very Good	6
EVAL21	83%	69%	76%	Very Good	7
GUCCE	79%	72%	76%	Very Good	8
EPOOL	87%	64%	76%	Very Good	9
OCHMU	84%	66%	75%	Very Good	10
DGDFM	83%	64%	74%	Very Good	11
IDAII	86%	61%	73%	Very Good	12
ESHIV	75%	72%	73%	Very Good	13
NCDKE	66%	74%	70%	Very Good	14
VTCAR	88%	51%	70%	Very Good	15
SUPCH	78%	61%	70%	Very Good	16
HREVA	78%	61%	70%	Very Good	17
ARCHE	71%	68%	69%	Very Good	18

MASTE	86%	49%	67%	Very Good	19
USCOV	79%	56%	67%	Very Good	20
FRCOH	81%	53%	67%	Good	21
COMME	69%	63%	66%	Good	22
DIGHP	71%	60%	65%	Good	23
HIVKIN	65%	63%	64%	Good	24
OCBFE	78%	45%	62%	Good	25
BUDGE	80%	43%	62%	Good	26
VOTTR	57%	60%	59%	Good	27
MVGCE	68%	38%	53%	Good	28
OCBPR	70%	35%	53%	Good	29
MAURT	52%	33%	42%	Good	30
EMRKS	45%	37%	41%	Fair	31

ANNEX VI: DETAILED UNEG RATING: GOOD

This annex contains meta-evaluative judgments about the portfolio of evaluations from 2017-2021 and SEU evaluation system using the United Nations Evaluation Group Norms and Standards Framework. Two summary tables report the ratings for all Norms and Standards. These tables are followed by detailed assessments. Each Norm and Standard assessment contains an abbreviated definition, description of actual performance, evidence sources for description and rating, performance rating, and, if applicable, standard-specific recommendations. This quality framework was not applied to each individual evaluation, but used ratings at the aggregated portfolio-level along with other sources of evidence to reach these evaluative conclusions about the SEU system and past 5 years of evaluation performance. The levels of performance quality are the same levels used in the PrgES and ALNAP Proforma frameworks (Excellent, Very Good, Good, Fair, and Poor). These ratings are not based on a predetermined ratio of met or not met standards and a systematic formula for synthesis, but on a generic qualitative rubric shared across all norms and standards as shared below and interpreted by the meta-evaluation team. Only two norms of the UNEG 38 norms and standards were deemed not applicable to the MSF/OCB/SEU context, which were Norm 1 about the Sustainable Development Goals and Norm 9 about national evaluation capacity. Standard ratings are based on average sub-standard ratings that are rounded to the nearest whole number and rounded up at half points. Ratings of norms, standards, and sub-standards were averaged with the following numerical codes (poor=1; fair=2; good=3; very good=4; excellent=5).

NORMS AND STANDARDS RUBRIC

Excellent	**Very Good**	**Good**	**Fair**	**Poor**
Norm or standard is completely manifest in actual performance and supported by strong evidence.	Norm or standard is mostly manifest in actual performance and supported by strong evidence.	Norm or standard is partially manifest in actual performance as indicated by sufficient evidence.	Norm or standard is partially to mostly not manifest in actual performance as indicated by sufficient evidence.	Norm or standard is mostly or completely not manifest in actual performance supported by strong evidence of or there is no evidence to support claims of meeting this standard and this lack of evidence is treated as lack of performance standard in action.

UNEG NORMS AND STANDARDS RATINGS

Table: UNEG Norms Ratings

Norm[76]	Definition Excerpt	Rating
Utility	"...there should be a clear intention to use [the evaluation] to inform decisions and actions"	VERY GOOD
Credibility	"Credibility is grounded on independence, impartiality and a rigorous methodology."	VERY GOOD
Independence	"...evaluators [should] be impartial and free from undue pressure throughout the evaluation process."	VERY GOOD
Impartiality	"The key elements of impartiality are objectivity, professional integrity and absence of bias."	FAIR
Ethics	"integrity and respect for [culture]...human rights... gender equality; and 'do no harm'"	GOOD
Transparency	"Evaluation products should be publicly accessible."	VERY GOOD
Human Rights and Gender Equality	"[integrate] principles of human rights and gender equality... into all stages of an evaluation."	FAIR
Professionalism	"Evaluations should be conducted with professionalism and integrity."	VERY GOOD
Enabling Environment	"an organizational culture that values evaluation as a basis for accountability, learning and evidence-based decision-making"	GOOD
Evaluation Policy	"clear explanation[s] of the purpose, concepts, rules and use of evaluation within the organization"	VERY GOOD
Responsibility for the Evaluation Function	"governing body [is] responsible for...independent, competent and adequately resourced evaluation [unit]"	VERY GOOD
Evaluation Use and Follow-up	"...promote evaluation use and follow-up, using an interactive process that involves all stakeholders."	POOR
OVERALL NORM RATING FOR EVALUATION PORTFOLIO and SYSTEM		GOOD

[76] Two Norms, *1. Agreed Principles, Goals, Targets* and *9. National Evaluation Capacities* were deemed not applicable to the SEU evaluation portfolio and evaluation system as defined. If included, Norm 1 would correspond to SEU's assessment of evaluation alignment with Strategic Orientations and Operational Priorities and Norm 9 would correspond to SEU's evaluation capacity development at the Operational Cell level.

Table: UNEG Standards Ratings

UNEG Standards Ratings for SEU Evaluation Portfolio and System	
Standard and Sub-standard	Rating
INSTITUTIONAL FRAMEWORK	GOOD
Institutional Framework for Evaluation	VERY GOOD
Evaluation Policy	VERY GOOD
Evaluation plan and reporting	VERY GOOD
Management response and Follow-up	POOR
Disclosure Policy	GOOD
MANAGEMENT OF THE EVALUATION FUNCTION	VERY GOOD
Head of Evaluation	VERY GOOD
Evaluation Guidelines	VERY GOOD
Responsiveness of Evaluation Function	VERY GOOD
EVALUATION COMPETENCIES	VERY GOOD
Competencies	VERY GOOD
Ethics	GOOD
CONDUCT OF EVALUATIONS	GOOD
Timeliness and Intentionality	GOOD
Evaluability Assessment	POOR
Terms of Reference	VERY GOOD
Scope and Objectives	GOOD
Methodology	GOOD
Stakeholder Engagement and Reference Groups	VERY GOOD
Human Rights-based Approach and Gender Mainstreaming	FAIR
Selection and Composition of Evaluation Teams	GOOD
Evaluation Report and Products	GOOD
Recommendations	GOOD
Communication and Dissemination	VERY GOOD
QUALITY	GOOD
Quality Assurance System	GOOD
Quality of the evaluation design	GOOD
Quality of the final stage of the evaluation	FAIR
OVERALL STANDARD RATING FOR EVAL PORTFOLIO and SYSTEM	GOOD

GENERAL NORMS FOR EVALUATION: GOOD

1. AGREED PRINCIPLES, GOALS, TARGETS: (NA)

Definition: "Within the United Nations system, it is the responsibility of evaluation managers and evaluators to uphold and promote, in their evaluation practice, the principles and values to which the United Nations is committed. In particular, they should respect, promote and contribute to the goals and targets set out in the 2030 Agenda for Sustainable Development."

SEU Description: NA.

Evidence Sources: NA.

Recommendation: NA.

2. UTILITY: VERY GOOD

Definition: "In commissioning and conducting an evaluation, there should be a clear intention to use the resulting analysis, conclusions or recommendations to inform decisions and actions. The utility of evaluation is manifest through its use in making relevant and timely contributions to organizational learning, informed decision-making processes and accountability for results. Evaluations could also be used to contribute beyond the organization by generating knowledge and empowering stakeholders."

SEU Description: The highest rating of all quality criteria in the PrgES framework for the SEU was Utility (Very Good). There is ample evidence to demonstrate that the SEU attends to utility at multiple points in the evaluation process, from a stand alone section in the ToR, to consultation group formation, and the actual scores and ratings from the 48 Utility indicators used across the past 5 years of evaluations. Utility according to this UNEG definition includes actual instrumental findings use, along with other forms of actual use such as process use and conceptual use. While utility for prospective use at the SEU is very good with strong evidence, actual evaluation use is weaker and with a weaker evidence base. This weaker rating is mostly reflected in the norm 14 on Evaluation Use and Follow-up. This particular Utility norm rating relies mostly on prospective use that may not have actually occurred, with one marginally weighted metric of actual use from the online survey, use satisfaction.

Evidence Sources: Use and Dissemination Plans; PrgES Utility Ratings; ALNAP Context and Report Ratings; Cost/Utility Metric (including use and dissemination satisfaction)

Recommendation: As an SEU and Steering Committee, consider what factors contribute to a significant differential in ratings between prospective use (utility) and actual use and influence (use and follow-up). It may be actual use and influence observations in this meta-evaluation are deflated and inadequately capture the true effect of evaluations, and or that utility measures are systematically inflated for some unaccounted threat to validity, or it may be a lack of coherence in terms of how the SEU supports intended users after the report has been written and delivered. At any rate, higher utility ratings suggest un or under-used and likely undervalued evaluations. Additional recommendations are offered for actual use in norm 14.

3. CREDIBILITY: VERY GOOD

Definition: "Evaluations must be credible. Credibility is grounded on independence, impartiality and a rigorous methodology. Key elements of credibility include transparent evaluation processes, inclusive approaches involving relevant stakeholders and robust quality assurance systems. Evaluation results (or findings) and recommendations are derived from — or informed by — the conscientious, explicit and judicious use of the best available, objective, reliable and valid data and by accurate quantitative and qualitative analysis of evidence. Credibility requires that evaluations are ethically conducted and managed by evaluators that exhibit professional and cultural competencies."

SEU Description: Evaluator credibility sub-criterion (PrgES U1) was rated "good" as well as the overall accuracy and propriety ratings. The SEU has policies and procedures that ensure independence and impartiality in external evaluations. The ALNAP methods rating for the portfolio was very good as evidenced by specific indicators from evaluation reports. Attention to professionalism and cultural competence was evident in our interactions with the head of unit and evaluation managers, as well as manifest in the evaluation artifacts.

Evidence Sources: PrgES Utility, Accuracy, Proprietary Ratings; ALNAP Methods Domain Ratings; SEU guidance documents

Recommendation: this credibility rating is one level shy of "Excellent". Marginal gains across most, if not all of the quality criteria would logically improve credibility ratings.

4. INDEPENDENCE: VERY GOOD

Definition: "Independence of evaluation is necessary for credibility, influences the ways in which an evaluation is used and allows evaluators to be impartial and free from undue pressure throughout the evaluation process. The independence of the evaluation function comprises two key aspects — behavioural independence and organizational independence. Behavioural independence entails the ability to evaluate without undue influence by any party. Evaluators must have the full freedom to conduct their evaluative work impartially, without the risk of negative effects on their career development, and must be able to freely express their assessment. The independence of the evaluation function underpins the free access to information that evaluators should have on the evaluation subject. Organizational independence requires that the central evaluation function is positioned independently from management functions, carries the responsibility of setting the evaluation agenda and is provided with adequate resources to conduct its work. Organizational independence also necessitates that evaluation managers have full discretion to directly submit evaluation reports to the appropriate level of decision-making and that they should report directly to an organization's governing body and/or the executive head. Independence is vested in the Evaluation Head to directly commission, produce, publish and disseminate duly quality-assured evaluation reports in the public domain without undue influence by any party."

SEU Description: The SEU received two excellents, one very good, and one good rating for the applicable sub-criteria for this independence norm. The SEU as a unit appears to be adequately resourced and given adequate autonomy to manage third-party independent external evaluations for OCB. Evaluation policies declare independence as a stated evaluation norm, and evidence in contractual language and evaluation reports supports this reality. There was no evidence of undue influence by management that limited the autonomy of the SEU to manage external evaluations. The commission and expenditure of this external independent meta-evaluation alone also weighs into this high rating.

Evidence Sources: PrgES Propriety 2, 5, 6, and Evaluation Accountability 3 Ratings; ALNAP 2.1 Rating; Evaluation Contracts; Evaluation Policies (framework, manifesto, roles and responsibilities)

Recommendation: SEU managers and head of unit need to reflect on whether they ever feel or have felt unduly influenced to deliver or modify evaluation results directly or indirectly, intentionally or unintentionally. While there may be no evidence of dependence to meta-evaluators, this was not explicitly asked of SEU members and could be a reality. Honest conversations as a unit or with the head of the unit that result in suggestions of any undue external influence should be reported to the steering committee with a plan for evaluation policy or procedure solutions. Consider adding an explicit clause in the evaluation contract that while MSF has intellectual rights to evaluation reports, external evaluators have final editorial authority in terms of evaluative content. Consider light touch independent reviews of evaluation inception reports and evaluation plans.

5. IMPARTIALITY: FAIR

Definition: "The key elements of impartiality are objectivity, professional integrity and absence of bias. The requirement for impartiality exists at all stages of the evaluation process, including planning an evaluation, formulating the mandate and scope, selecting the evaluation team, providing access to stakeholders, conducting the evaluation and formulating findings and recommendations. Evaluators need to be impartial, implying that evaluation team members must not have been (or expect to be in the near future) directly responsible for the policy setting, design or management of the evaluation subject."

SEU Description: Though housed within the operations department, members of the SEU are not involved with the policy-setting, design, or management of the objects of evaluation. Many of the SEU managers were hired internally, which increases their credibility and capacity to understand the institutional context, but despite this organizational familiarity, there is no evidence to suggest their experience prevents them from impartial management. For external evaluators, there were multiple instances of evaluation consultants being hired from within the MSF talent pool, meaning evaluators who may have first and foremost been subject-matter experts to the specific evaluation subject with institutional experience supported by some degree of sufficient evaluation know-how. Members of the MSF movement place high-value on evaluator contextual knowledge of the distinguishing values, principles, and behavioral commitments along with systematic operational constraints. However, this positionality may present a limitation in impartiality, especially if undisclosed in final reports with no recognition of or plans to mitigate potential biases. The relevant ALNAP standard that pertained to this norm systematically received "poor" ratings due to little to no reports of the nature and make up of the evaluations team nor how these compositions may or may not bias evaluation processes. Further, inconsistent attestations of conflicts of interest in inception reports and final reports also lowered the score for the respective PrgES sub-criterion. Low reliability ratings also influenced this rating in terms of mitigating biases. However, multiple conversations with the head of the SEU included reflexivity to mitigate any undue biasing of the process with prematurely disclosed judgments or personal opinions.

Evidence Sources: PrgES U8, F3, Proprietary Ratings, A3, A2; ALNAP 1.2 Rating; Evaluation Contracts; Evaluation Policies (framework, manifesto, roles and responsibilities); Participant observation

Recommendations: mandate conflict of interest statements in the evaluation report template; mandate a section within the inception report methods section that invites evaluator teams to speak to the evaluator or team composition and how their lived experience may constrain or enable more accurate evaluative conclusions; ensure managers invite evaluation teams to speak to how they plan on ensuring reliability in the inception report; hold discussions with the SEU steering committee about the benefits and drawbacks of SEUs policy (tacit or explicit) toward hiring former or current MSF movement members; make any policy decisions transparent.

6. ETHICS: GOOD

Definition: "Evaluation must be conducted with the highest standards of integrity and respect for the beliefs, manners and customs of the social and cultural environment; for human rights and gender equality; and for the 'do no harm' principle for humanitarian assistance. Evaluators must respect the rights of institutions and individuals to provide information in confidence, must ensure that sensitive data is protected and that it cannot be traced to its source and must validate statements made in the report with those who provided the relevant information. Evaluators should obtain informed consent for the use of private information from those who provide it. When evidence of wrongdoing is uncovered, it must be reported discreetly to a competent body (such as the relevant office of audit or investigation)."

SEU Description: The SEU has an excellent evaluation policy document that articulates expectations with ethical evaluation practice. The meta-evaluators and other primary evaluators were expected to read and sign the document since its recent publication. An ethics lens was a stated emphasis of this meta-evaluation and it is clear from MSF-wide documents that ethical practice is a strong value, as well as support for generally accepted humanitarian principles, not the least of which, "do no harm." Many of these policies are inscribed in clauses that pertain to ethics in the evaluation contract, including minimum behavioral standards. With that aspiration, this portfolio of evaluations received a very good rating across evaluations for the responsive and inclusive orientation criterion, and mixed results with a 50% score and subsequent "good" rating for human rights and respect, with informed consent practices very good, but not excellent in all cases. The overall Propriety rating from the PrgES framework was rated as "good". Only a third of the managed evaluations attended to gender in analysis, and about a third were missing cross-cutting analyses attending to vulnerable populations and protection, despite these being two strong values for MSF. Information from interviews and surveys suggest practices are mixed to weak in terms of downward accountability to communities for evaluation findings. Information management received a fair rating, and at least one call-out from a prior evaluator as an area of concern or improvement for the SEU. While possibly a higher bar than a cosmopolitan notion of ethics, absent across most evaluations were considerations of culturally responsive/specific and equitable evaluation practice.

Evidence Sources: PrgES Proprietary Ratings; ALNAP Cross-cutting Issue (4.4) Ratings; Evaluation Contracts; Evaluation Policies (framework, manifesto, roles and responsibilities)

Recommendation: review and assess the SEU performance against the UNEG Ethical Guidelines for Evaluation framework, including associated role-specific checklists. Identify any policy gaps in comparing that framework with the SEU ethical evaluation policy document.The SEU ethical policy document was informed by this UNEG Norms and Standards framework and the UNEG Ethical Guidelines for Evaluation is a different and more detailed document on ethical evaluation. Consider templating the cross-cutting questions of vulnerable populations, protection, and gender equity in evaluation report templates (or a similar marginalized group analysis), and where such analysis is not applicable, having evaluation teams make such declarations with sufficient explanation.

7. TRANSPARENCY: VERY GOOD

Definition: "Transparency is an essential element of evaluation that establishes trust and builds confidence, enhances stakeholder ownership and increases public accountability. Evaluation products should be publicly accessible."

SEU Description: The SEU publicly shares almost all of the evaluations it manages. Aside from a couple evaluations that had insufficient final report quality ratings, and instances where there was no final written report, there was only one case of the evaluation portfolio that was not shared due to conflict sensitivity concerns. Most of the evaluations are widely disseminated within the MSF movement and among appropriate right-to-know audiences at the cell or country level. Checkpoints that scored low pertained to equipping all right-to-know audiences and participants with information about evaluation policies as well as information about monetary sources, missing information about the nature of the evaluator selection process, as well as information about fairness and transparency in allocating finite evaluation resources to meet multiple evaluation participant and right to know audience needs.

Evidence Sources: PrgES P4, P5, A8, Evaluation Accountability Ratings; ALNAP Sections 1 (TOR) and 5 (Report); Contracts; Use and Dissemination Plans

Recommendation: consider and enact the appropriate policy or procedural changes to mitigate low scores for specific indicators of sub-criteria related to transparency in the PrgES and ALNAP proforma.

8. HUMAN RIGHTS AND GENDER EQUALITY: FAIR

Definition: "The universally recognized values and principles of human rights and gender equality need to be integrated into all stages of an evaluation. It is the responsibility of evaluators and evaluation managers to ensure that these values are respected, addressed and promoted, underpinning the commitment to the principle of 'no-one left behind'."

SEU Description: While the SEU does not espouse an explicitly human rights-based or -centered approach to evaluation, attention to basic human rights is attended to in the SEU ethical policy document, which is predominantly sourced from the UNEG Norms and Standards document. While the SEU portfolio is rated "good" for the Propriety criterion, there are sub-criteria and sub-domains that are not as well rated. Only a third of the managed evaluations across the portfolio attended to gender in analysis, and about a third were missing cross-cutting analyses attending to vulnerable populations and protection, despite these being two strong values for MSF. Most evaluation cases did not produce evidence that groups traditionally excluded from or hindered by evaluation processes were sought out.

Evidence Sources: ALNAP 1.2, 2.5, 4.4 Gender Standard; PrgES P3; Evaluation Contracts; Evaluation Policies (framework, manifesto, roles and responsibilities)

Recommendation: Consider referring to UNEG guidance and technical scorecard for ideas on mainstreaming human rights and gender equity in evaluation practice. Additionally, following comments with the head of unit about internal political will about MSF's attention to institutional complicity with global structures of exploitation and colonization, following the "Do No Harm" humanitarian principle, commission an internal or external transversal desk report on transformative evaluation approaches, and specifically Culturally Responsive and Equitable Evaluation (CREE) models and evaluation policies and the potential benefit these practices provide OCB in being a leader in aligning practices with aspirations with their evaluation function. A possible entry-point to this conversation could be comparing the analysis of the Cultural Reading of the 2nd Edition of the Program Evaluation Standards with the PrgES checklist that was used for this meta-evauation. An additional source of resources might be the Funder and Evaluator Affinity Network Call to Action series that presents concrete issues related to equity from common evaluation policies and practices. There is also this related document: Righting Systemic Wrongs Organizational Self-Assessment.

9. NATIONAL EVALUATION CAPACITIES: NA

Definition: "The effective use of evaluation can make valuable contributions to accountability and learning and thereby justify actions to strengthen national evaluation capacities. In line with General

Assembly resolution A/RES/69/237 on building capacity for the evaluation of development activities at the country level, national evaluation capacities should be supported upon the request of Member States."

SEU Description: NA (explain the next relevant consideration within SEU context, but ultimately why this is not being included).

Evidence Sources: NA.

Recommendation: NA.

10. PROFESSIONALISM: VERY GOOD

Definition: "Evaluations should be conducted with professionalism and integrity. Professionalism should contribute towards the credibility of evaluators, evaluation managers and evaluation heads, as well as the evaluation function. Key aspects include access to knowledge; education and training; adherence to ethics and to these norms and standards; utilization of evaluation competencies; and recognition of knowledge, skills and experience. This should be supported by an enabling environment, institutional structures and adequate resources."

SEU Description: This meta-evaluation did not explicitly draw on any evaluator competencies framework, nor were the SEU personnel the primary object of meta-evaluation. However, the SEU is staffed by knowledgeable, skilled, and ethical professionals. Our participant observation and experience in being managed by the SEU largely informs our eventual rating of this norm. The SEU is aware of and attempts to embody important evaluator competencies, and references the American Evaluation Association evaluator competencies in their policy documents and purports to follow a competency-based selection process. Evaluation policy and practice points are debated, codified, and enacted in the unit through supportive evaluation management and accompaniment. Utility sub-criterion about credibility scores "good" along with the Propriety criterion. The Feasibility criterion is 2 percentage points shy of the "very good" cut score. Overall portfolio ratings suggest the SEU has been able to attract and contract with external evaluators with a high degree of professionalism.

Evidence Sources: Participant Observation; PrgES U1, Feasibility, Propriety

Recommendation: consider small changes to regular SEU meetings to foster ongoing professional development moments such as having a rotating schedule of managers doing a brief show and tell moment at the start of meetings highlighting a specific evaluator competency from the AEA competencies framework or guiding principles, the Program Evaluation Standards, UNEG Norms and Standards, or MSF-specific sensitizing principles. Consider adding individual professional development plans for evaluation managers to annual plans where managers and the head of unit take individual self-assessments of evaluator competencies and make goals and plans to improve in core competencies with regular informal or formal check-ins about professional progress.

11. ENABLING ENVIRONMENT: GOOD

Definition: "Evaluation requires an enabling environment that includes an organizational culture that values evaluation as a basis for accountability, learning and evidence-based decision-making; a firm commitment from organizational leadership to use, publicize and follow up on evaluation outcomes; and recognition of evaluation as a key corporate function for achieving results and public accountability. Creating an enabling environment also entails providing predictable and adequate resources to the evaluation function."

SEU Description: The SEU recently passed its 10-year mark in terms of being organized with a mandate to support operations at OCB. There are many factors that inform rating this norm, which include policy documents, interviews with other support teams, knowledge of key evaluation support mechanisms such as the steering committee, mandated consultation groups, standard operating procedures before, during, and after evaluations, key annual events such as evaluation days, and reports to OCB board, webinars, transversal learning inquiries and reports. Many of these artifacts, events, practices, and values combine into an enabling environment that is broadly supportive of the evaluative function, not the least of which are anecdotes of cells and units within OCB that have come to expect a level of quality and rigor with SEU products and processes, which translates to a healthy reputation and increased demand for evaluation work. Also, at least a recently updated evaluation policy that all programs are to be evaluated, unless there is good reason not to. These are promising aspects of the overall evaluation environment at SEU. Annual reports also provide insights that challenges the SEU faced in 2018 and 2019 have abated to some extent. However, there are some indications that these systems are still not being leveraged to their full potential in terms of actual use and influence of evaluation findings and processes. Insights from managers suggest there is general room for improvement in monitoring and evaluation systems, possibly in regular data collection at the cell and program level. Comments from evaluators indicate consistent lack of documented program theory in terms of logical frameworks, logic models, or theories of change, which limited evaluation activities. Limited evidence about use of evaluations, few management responses, a view held by the SEU steering committee that use and influence of evaluation by operations (what happens after dissemination) is mostly outside the remit of the SEU, assumptions about the lack of the necessity of the evaluation function and operations during a global pandemic from senior leadership suggest there is an opportunity to improve the enabling environment. Concrete gains can be made in cells and operations teams improving evaluability through minor improvements to program design and performance monitoring and in the role of evaluation commissioners and evaluation intended users in ensuring intended use is realized after external evaluation consultants leave the picture.

Evidence Sources: SEU Guiding documents; Transversal learning and Evaluation Day Material; Interviews with Head of SEU; PrgES and ALNAP Scores and Ratings; Evaluation Budgets

Recommendation: celebrate the successes of the past 5 years of evaluation work by highlighting the strengths of the evaluation portfolio and evaluation system. Where there are sub-standard criteria and sub-criteria, highlight the benefit of having detailed maps of quality in terms of where adjustments can be made to make improvements and deliver more valuable evaluations to operations. Use the moment of being evaluated to connect and empathize with those who have been or might be evaluated to address concerns they may have about the potential benefit of evaluation. More concretely, if OCB has a policy that every program is to be evaluated, assemble a checklist that articulates minimum viable standards for programs to integrate into their management and monitoring to improve the evaluability of those programs. Consider collaborative evaluability assessments with cells, where cells co-evaluate the extent to which their operations could be evaluated, which could improve the enabling environment through concrete evaluability gains and help cells realize their own demand in terms of the potential value evaluations could add to operations.

12. EVALUATION POLICY: VERY GOOD

Definition: “Every organization should establish an explicit evaluation policy. Taking into account the specificities of the organization’s requirements, the evaluation policy should include a clear explanation of the purpose, concepts, rules and use of evaluation within the organization; the institutional framework and roles and responsibilities; measures to safeguard evaluation independence and public accountability; benchmarks for financing the evaluation function that are commensurate with the size and function of the organization; measures to ensure the quality and the use of evaluations and post-evaluation follow-up; a framework for decentralized evaluations, where applicable; and provision for periodic peer review or external assessment. The evaluation policy should be approved by the governing body and/ or the executive head to ensure it has a formally recognized status at the highest levels of the organization. References to evaluators in the policy should encompass staff of the evaluation function as well as evaluation consultants."

SEU Description: The SEU has a decent array of evaluation guidance documents that articulate and explain in detail the values, principles, and procedures that constitute good evaluation practice at OCB. Together, this evaluation policy contextualizes many industry standards to institutional and operational settings. Roles and responsibilities are delineated, as well as specific actions for each role by the SEU-specific evaluation stages. Participant observation and review of artifacts suggest these policies are being enacted for the most part consistently across evaluation cases. A systematic policy around external assessment or peer review (meta-evaluation) is absent. Additionally, at the outset of the meta-evaluation inquiry, the SEU indicated no quality framework was developed to serve as criteria and standards for the review and that one would need to be developed. The meta-evaluators were surprised to encounter the breadth of quality statements across the policy documents and wondered why these policies and statements were not foregrounded more fully as the basis for meta-evaluative claims.

Evidence Sources: SEU Guiding Documents; SEU annual goals

Recommendation: The SEU can be more explicit in owning components of policy documents as the basis for quality assessment and meta-evaluation at OCB. Further, we suggest consolidating material across policy documents into one coherent policy document that houses values, principles, criteria, and standards of good evaluation practice. Further, any eventual consolidated quality framework might consider what principles or practices, either presently included or absent from SEU documents, are exclusive to the conduct of evaluating medical humanitarian interventions and need to be highlighted for external evaluation consultants to consider.

13. RESPONSIBILITY FOR THE EVALUATION FUNCTION: VERY GOOD

Definition: "An organization's governing body and/or its executive head are responsible for ensuring the establishment of a duly independent, competent and adequately resourced evaluation function to serve its governance and management needs. The evaluation budget should be commensurate to the size and function of the organization. The governing body and/or the executive head are responsible for appointing a professionally competent head of evaluation and for fostering an enabling environment that allows the head of evaluation to plan, design, manage and conduct evaluation activities in alignment with the UNEG Norms and Standards for Evaluation. The governing body and/ or the executive head are responsible for ensuring that evaluators, evaluation managers and the head of the evaluation function have the freedom to conduct their work without risking their career development. Management of the human and financial resources allocated to evaluation should lie with the head of evaluation in order to ensure that the evaluation function is staffed by professionals with evaluation competencies in line with the UNEG Competency Framework. Where a decentralized evaluation function exists, the central evaluation function is responsible for establishing a framework that provides guidance, quality assurance, technical assistance and professionalization support."

SEU Description: The SEU is a centralized evaluation unit—for Operational Centre Brussels—that does not conduct evaluations, but manages the conduct of external evaluation consultants. The unit is run by a professional, skilled, and ethical head of unit that has been leading efforts to consolidate evaluation policy, increase institutional reputation, and promote a broad culture of evaluation among primary evaluation users and prospective users. Sorting evaluation cases by year reveals annual increases in evaluation quality over the past five years. The head of unit oversees a team of trained and capable managers, equipped with different backgrounds, strengths, and capacities. No evidence exists to suggest that the head of unit or managers are unable to fulfill their responsibilities by limitations caused by their organizational structure or position within OCB. Analysis of survey data suggests that managers have a very detailed and thorough understanding of their responsibility in ensuring evaluation quality, both for specific cases and in general, at all stages, but especially in the scoping, preparation, and inception stages. Evaluation commissioners cite more responsibility in scoping and dissemination and use stages. This recognition coupled with performance data on use and follow-up suggest greater accountability measures may need to be in place for commissioners to ensure the potential value of high-quality evaluations are being realized. Limited project contact data about roles and responsibilities suggest greater importance on data collection and analysis in terms of serving as a link between evaluation and operational teams. The evaluation coordinator did not respond to the question about roles and responsibilities in the survey by design, but interview data suggests a clear understanding of the role. The meta-evaluation team is aware of efforts to encourage more participation from internal evaluation clients within cells to support administrative functions of evaluation commissioning to free up SEU administrators to focus on promoting use and dissemination efforts as well as transversal learning.

Evidence Sources: SEU Roles and Responsibilities; Survey data about responsibility of roles; Interviews with Head of SEU; Interviews with SEU managers; Other SEU policy documents.

Recommendation: With a recent loss of an evaluation manager, and increases in evaluation demand, it is likely the SEU and OCB would stand to benefit from hiring additional manager(s). Further, the SEU may consider designating SEU focal points for key cross-cutting evaluation functions and stages, playing to the strengths of existing managers. For example, while all managers might be responsible for a portfolio of multiple open evaluations at any one time, one manager might be charged with ensuring sufficient evaluability processes have occurred across the unit and another may oversee ensuring use and follow-up are not going unaddressed. Further, it is evident that more work is needed to equip and empower evaluation commissioners and intended users at the cell-level to adequately carry forward evaluation findings and recommendations. While cost-utility analysis has been conducted, total and average evaluation budgets relative to total evaluation object program budgets have not been compared to arrive at any statements about the relative appropriateness of evaluation budgets to operational budgets. UNEG guidance suggests 3% - 5% of total program budget should be allocated to the evaluation function. Consider comparing average total expenses for OCB for the years of the evaluations with the total annual evaluation budget to see how the evaluation function budget relative to total expenses compares to this or other industry standard recommendations.

14. EVALUATION USE AND FOLLOW-UP: POOR

Definition: “Organizations should promote evaluation use and follow-up, using an interactive process that involves all stakeholders. Evaluation requires an explicit response by the governing authorities and/or management addressed by its recommendations that clearly states responsibilities and accountabilities. Management should integrate evaluation results and recommendations into its policies and programs. The implementation of evaluation recommendations should be systematically

followed up. A periodic report on the status of the implementation of the evaluation recommendations should be presented to the governing bodies and/or the head of the organization."
SEU Description: There is strong evidence that Evaluation Use and Follow-up are "Fair" at OCB. This rating is the result of splitting the difference between the two constructs the UNEG framework combines for this norm–Follow-up and Use. These should ideally be two distinct Norms that receive their own measures and ratings as an evaluation can be used without it being followed up, or followed up only to find out there was no use. With that, limited management responses and follow-up documentation resulted in a "Poor" Follow-up rating. Detailed analyses about evaluation use and use outcomes in the mEQ3 findings section reveal evaluation use at OCB is "Good."

Evidence Sources: Use and Dissemination Plans; Survey response data on evaluation use and consequences; Management Response Documents; Use Satisfaction Ratings; Use and Influence Ratings
Recommendation: The meta-evaluation team acknowledges mere application of evaluation recommendations is not the only or most important marker of quality in terms of use and follow-up. It may be that recommendations from external evaluators are poorly supported, not feasible, inappropriate, untethered to evaluative conclusions and findings, and not culturally responsive or specific. Differences of opinion remain about the role of evaluators in making recommendations and the reviewers appreciate the SEU's position about the collaborative nature of recommendation generation, per policy documents. That being said, the MSF context does seem to place high value on "findings use" or being able to make decisions and take action from quality recommendations that follow logically from sound evaluative conclusions. The value placed on this type of intended use of evaluation and the actual performance of this follow-up function suggest a major disconnect in aspiration and reality. Some significant remediation plan for evaluation follow-up is needed from the SEU and SEU steering committee. While it is the view of the meta-evaluators that the responsibility for evaluation use rests across many roles, the evaluation commissioner is primarily responsible for evaluation use and the SEU is primarily responsible for evaluation use follow up. Consider developing a recommendation rubric that defines the dimensions of a quality recommendation according to the SEU and OCB. Invite evaluators to use this rubric when creating or co-creating recommendations with evaluation participants. Invite consultation group members and especially commissioners to use this in their follow-up and response. Finally, consider having both commissioners draft a functional management response in collaboration with the SEU head of unit that is published as an annex of the meta-evaluation to increase transparency and credibility in terms of committing to areas that the meta-evaluation identifies as needing improvement. Write evaluation policy and enact procedures that duplicate this practice for primary evaluations, that is, include a management response as a public annex by default. Create a rubric of general evaluation quality in that management response template that managers can rate so the content of their responses is more transparent.

STANDARDS[77] FOR EVALUATION: GOOD

For ease of readability, detailed definitions, sub-criteria, and specific indicators of the following standards have been omitted in the definition sections of this document, but were referred to for comparison to arrive at performance ratings. More detailed explanations of these standards can be

[77] Standard ratings are based on average sub-ratings that are rounded to the nearest whole number and rounded up at half points. Ratings of standards and sub-standards were averaged with the following numerical codes (poor=1; fair=2; good=3; very good=4; excellent=5).

found here. Some standards are duplicative of Norms and, where applicable, descriptions and recommendations may have already been reported above and indicated as such.

1. INSTITUTIONAL FRAMEWORK: GOOD

1. Institutional Framework for evaluation: Very Good

Definition: "The organization should have an adequate institutional framework for the effective management of its evaluation function."

SEU Description: all available evidence suggests the SEU has an adequate support structure in terms of steering committee, support from board, executive leadership, adequate human resources. Data on use and influence suggest improvements could be made in extending integration of evaluation function with management decisions at cell and operational centre level. No formal analysis of SEU budget relative to total operational budget has been conducted, but such analysis would provide insight into the relative appropriateness of the SEU framework for OCB. Continued improvements in evaluation quality, and especially evaluation value (through improvements of the use and follow-up) will strengthen OCB management's understanding and support for the evaluation function to contributing to the effectiveness of the operational centre.

Evidence Sources: SEU Evaluation Framework Document; Key informant interviews

Recommendation: Consider the head of unit analyzing percent of SEU budget relative to total operational budget and report to the steering committee and OCB board the degree of appropriateness of financial and human resource allocations. Ensure meta-evaluation conclusions and recommendations are weighed, prioritized, and translated into a use plan for evaluation system improvements.

2. Evaluation policy: Very Good

Definition: "Organizations should establish an evaluation policy that is periodically reviewed and updated in order to support the evaluation function's increased adherence to the UNEG Norms and Standards for Evaluation."

SEU Description: The SEU has a robust collection of policy documents. These could be disseminated widely as is, or consolidated further with more refinement given to what quality means to the SEU with any updated ideas following the meta-evaluation.

Evidence Sources: SEU policy documents

Recommendation: Consider consolidating evaluation policy documents, or those that could be disseminated outside of the SEU and those that may only be relevant for the SEU internally.

3. Evaluation plan and reporting: Very Good

Definition: "Evaluations should have a mechanism to inform the governing body and/or management on the evaluation plan and on the progress made in plan implementation."

SEU Description: The SEU has sufficient mechanisms to ensure management has detailed evaluation plans from the length inception stage. Inception reports were mostly detailed and adequate as design documents as well as quasi-project management charters. The SEU has adequate standardized reporting templates for inception reports and final reports.

Evidence Sources: SEU Evaluation Stages Document; PrgES U7, A8, Feasibility Ratings

Recommendation: Consider updating inception and final report templates to bid evaluation teams to attend to systematically unaddressed quality standards deemed important from the meta-evaluation product-oriented standards in the PrgES and ALNAP proforma checklists. Also, consider refining guidance around the expected information in evaluation matrices. There was variation in the quality of information in these design and planning tools across cases.

4. Management response and follow-up: Poor

Definition: "The organization should ensure that appropriate mechanisms are in place to ensure that management responds to evaluation recommendations. The mechanisms should outline concrete actions to be undertaken in the management response and in the follow-up to recommendation implementation."
SEU Description: see the detailed description under Norm 14
Evidence Sources: Management response documents; Survey data
Recommendation: see the detailed list of possible recommendations under Norm 14.

5. Disclosure policy: Good

Definition: "The organization should have an explicit disclosure policy for evaluations. To bolster the organization's public accountability, key evaluation products (including annual reports, evaluation plans, terms of reference, evaluation reports and management responses) should be publicly accessible."
SEU Description: The SEU has a good track record of publishing final evaluation reports. Some of these include the original terms of reference.
Evidence Sources: Evaluation Contracts; SEU Policy documents
Recommendation: consider if there are any other evaluation documents that could be shared publicly to increase accountability and credibility. We are aware some reports are not shared publicly, such as the inception report which often contains personally identifiable information of key informants, but these could be redacted if there is a perceived benefit of including this document along with ToRs or any other documents. One document to be shared that we believe could help the SEU and OCB rise to a new level of accountability is to share the management response to the evaluation as an annex in each evaluation report.

2. MANAGEMENT OF THE EVALUATION FUNCTION: VERY GOOD

1. Head of evaluation: Very Good

Definition: "The head of evaluation has the primary responsibility for ensuring that UNEG Norms and Standards for Evaluation are upheld, that the evaluation function is fully operational and duly independent, and that evaluation work is conducted according to the highest professional standards."
SEU Description: this meta-evaluation is not a direct personnel evaluation. However, it would be disingenuous to disassociate evaluation portfolio and system-level ratings from the management of the evaluation function, both the good and the bad. That being said, through interviews and multiple touch-points through the balance of this meta-evaluation, our independent judgment of the head of unit is that they are duly independent and likely doing more than most organizational evaluation units in the international aid industry for organizations of this size and nature to ensure fidelity to the highest professional standards, based on a cursory scan of publicly available meta-evaluations on the ALNAP Help Library.
Evidence Sources: Interviews with Head of Unit; Participant Observation

Recommendation: to the SEU steering committee and OCB operational senior management: support the head of unit in post-meta-evaluation use and dissemination plan, in particular in the follow-up of any remediation action plans.

2. Evaluation guidelines: Very Good

Definition: "The head of evaluation is responsible for ensuring the provision of appropriate evaluation guidelines."

SEU Description: See Norm 12 and standard 2.2.

Evidence Sources: SEU Guidance Documents

Recommendation: Consider consolidating various policy documents such as manifesto, ethical guidelines, and framework into one policy document or manifesto that articulates in one location these various aspects of evaluation purpose and quality; for ease of interpretation and application, consider framing quality dimensions and values as effectiveness principles, or values with verbs, such as the "Ask the right questions" which are prescriptive, easy to remember, and useful. Framing policy as principles versus rules chimes with the SEU's manifesto intent of providing a map for navigating complexity. Refer to the Principles-focused Evaluation Model for further guidance; consider use of a SEU created checklist or checklists for specific decision gates in evaluation lifecycle; consider developing and promoting the use of rubrics to encourage stronger evaluative reasoning; strongly suggest revising evaluation inception and final report template following consensus of most important areas for improvement. Process-oriented standards can be integrated into the IR template and product-oriented standards into the final report template.

3. Responsiveness of the evaluation function: Very Good

Definition: "The head of evaluation should provide global leadership, standard setting and oversight of the evaluation function in order to ensure that it dynamically adapts to new developments and changing internal and external needs."

SEU Description: The total Proprietary score is 66%, one percentage point away from a "Very Good" rating. This criterion is the most relevant domain associated with responsiveness. Additional evidence compliments this score. Time-series data shows increases in quality over the years. Evidence from contract amendments show responsiveness to unforeseen needs in terms of budget and timeline. A policy of regular communication between evaluation teams and managers, continual reflection and refinement of SEU practices are all indicative of responsiveness and adaptation.

Evidence Sources: Overall rating data; Survey data

Recommendation: An area for improvement is expanding consultation groups or reference groups to include individuals and groups who have been traditionally excluded or hindered from evaluation processes. Further, aside from the project management responsiveness, the SEU has an opportunity to lead out in increasing the cultural responsiveness and equity of evaluation practices, as mentioned in the ethics and human rights and gender equity norms. This is a broader interpretation of responsiveness, but in line with new developments in the evaluation field and internal and external needs for more equitable evaluation practices and policies generally.

3. EVALUATION COMPETENCIES: VERY GOOD

1. Competencies: Very Good

Definition: "Individuals engaged in designing, conducting and managing evaluation activities should possess the core competencies required for their role in the evaluation process."

SEU Description: As mentioned previously, this meta-evaluation did not utilize a comprehensive evaluation competency framework by design. This was not a personnel evaluation of evaluation managers, nor of prior evaluation consultants. That being said, internal survey respondents across all roles acknowledged responsibility for evaluation quality and success. This naturally includes evaluation competencies most importantly from evaluation consultants and managers. Synthesis of the evaluation case scores, ratings, and ranks across the PrgES and ALNAP frameworks, along with participant observation indicate high-degree of evaluation competency, in the majority of evaluation consultants, and in all of evaluation managers and the head of unit. Reviews of SEU policy documents and interviews with managers reveal sufficient knowledge of evaluation functions and awareness of core issues in evaluation practice. There is mention of an evaluation competency-based selection process for hiring external evaluators in the Manifesto.

Evidence Sources: PrgES and ALNAP Ratings; Participant Observation

Recommendation: Considering the SEU description, outside of formal certification or credentialing, evaluation competency assessments in professional settings are likely most useful when self-direct and used to design and plan for tailored professional development. If not already a feature of SEU continuous quality improvement, consider establishing a unit-wide practice of regular self-assessment and goal setting for professional development of evaluation competencies. Consider reviewing the competency-based selection process to see if the approach is yielding desired results. This meta-evaluation did not investigate evaluator selection tools such as score-cards or competency-based selection rubrics.

2. Ethics: Good

Definition: "All those engaged in designing, conducting and managing evaluations should conform to agreed ethical standards in order to ensure overall credibility and the responsible use of power and resources."

SEU Description: see detailed description in Norm 6 above.

Evidence Sources: Evaluation Contracts; Evaluation Ethical Guidelines; Participant Observation

Recommendation: see recommendation in the Norm 6 section above.

4. CONDUCT OF EVALUATIONS: GOOD

1. Timeliness and intentionality: Good

Definition: "Evaluations should be designed to ensure that they provide timely, valid and reliable information that will be relevant to the subject being assessed and should clearly identify the underlying intentionality."

SEU Description: Of all evaluations, 61% met standards for adequately indicating the rationale of the timing of the evaluation in the terms of reference. Of all evaluations 70% were mid-term (SEU's term for formative evaluation) indicating evaluations designed to provide useful information for operational management. 97% of evaluation ToRs adequately indicated purpose, objectives, and focus of evaluation, and 71% adequately defined intended users and intended uses in the terms of reference. The average duration of evaluations were 5.86 months, though no in-depth analysis into planned versus actual timelines were conducted. Deviations in timeline should not automatically be viewed as sub-standard performance with timeliness criterion, as other factors such as responsiveness and appropriateness to contingencies may need to be considered. Reports of dissatisfaction about evaluation duration were common from head of unit and managers from interviews, and from commissioners in survey data.

Evidence Sources: PrgES Utility and Accuracy Ratings; ALNAP Section 1 (ToR); Descriptive attribute data for evaluation cases

Recommendation: Tightening protocols around evaluability of program data, as well as focusing, limiting the number, and prioritizing evaluation questions may present improvements to timeliness of evaluations.

2. Evaluability assessment: Poor

Definition: "An assessment of evaluability should be undertaken as an initial step to increase the likelihood that an evaluation will provide timely and credible information for decision-making."

SEU Description: evidence of evaluation management artifacts such as pre-natal, stakeholder analysis document, evaluation checklists, and situational assessments suggest there have been attempts to determine and address evaluability. These limited documents were not shared with the meta-evaluation team, likely due to how infrequent they were used or that they were planned for but not developed. Scoping questions were shared for 8 of 31 evaluations. Interview data with the head of unit indicate significant insufficiency in terms of procedures related to what is known as evaluability assessments. Additional evidence such as survey data from evaluators, qualitative analysis of limitation sections, and actual performance of insufficient answers to scoped evaluation questions, mostly about outcomes and impact, suggest evaluability assessment is at best informal and inconsistent, and at worst non-existent. One formal evaluation policy, "Consider the evaluability of the project" exists as an operational principle under the "Method" domain in the evaluation manifesto. Seven of the 31 evaluations were awarded met standards for evaluability based on a generous interpretation of presence of scoping question documents. This standard is top two for most potential for improvement (the other being use and follow-up).

Evidence Sources: SEU management documents; PrgES Feasibility Standard 2, checkpoint 1 ratings; ALNAP Section 1 and 4.3 ratings; KIIs

Recommendation: Identify which existing procedures can be modified to address more fully the evaluability question. Use existing or create an SEU-specific evaluability checklist to be integrated to the scoping stage of the evaluation process and conducted collaboratively with program teams. Identify go/no go standards or thresholds associated with each chosen dimension of evaluability, which may include some action in-between go/no go that augments the nature of the evaluation exercise if the evaluation is still deemed to be a net positive.

3. Terms of reference: Very Good

Definition: "The terms of reference should provide the evaluation purpose, scope, design and plan."

SEU Description: The evaluation portfolio scored "Very Good" for section 1 of the ALNAP Proforma dedicated to assessing Terms of References. The lowest standard was related to having more information about the evaluator selection process. Two other standards related to evaluation timing and intended use could be improved in some instances where information was less than adequate. There is some variation, and some exemplar terms of reference from the portfolio, especially from recent evaluations, that can be used as models for future terms of reference.

Evidence Sources: ALNAP Section 1

Recommendation: with the assumption that terms of references and requests for proposals are the best approach for contracting evaluation consultants, the SEU is administering this function right. However, this standard is an instance where quality standards may be at odds. For instance, recent work has highlighted the potential limitations of the request for proposal process for equitable evaluation practice. In this, the SEU has an opportunity to ask if not only are they doing things right, but are they doing the right things, and do these existing generally accepted quality frameworks embody the values the SEU wants to promote. The meta-evaluation team recommends reviewing this specific FEAN Call to Action document: Evaluation is SO White: Systemic Wrongs Reinforced by Common Practices and How to Start Righting Them, in particular, but not limited to, the section about request for proposals and alternatives.

4. Evaluation scope and objectives: Good

Definition: "Evaluation scope and objectives should follow from the evaluation purpose and should be realistic and achievable in light of resources available and the information that can be collected."

SEU Description: the SEU spends a considerable amount of time in the inception stage of the evaluation process, and likely less time in the scoping phase of the evaluation process. Artifacts indicate most evaluations have clarity around purpose, scope, and objectives expressed as intended use, but these can be improved still. Of slight concern to the meta-evaluation team is the number of evaluation questions. The average number of questions prescribed across ToRs is roughly 15, which we understand encompass main questions and sub-questions. Some evaluations had as much as 29 questions prescribed and some with no prescribed evaluation questions. Though no statistical analysis was conducted, a rudimentary visual analysis of overall scores and question number does not reveal compelling trends with the number of evaluation questions prescribed and final scores, but it is likely the total average of evaluation questions can be reduced and their subsequent investigation improved. Some objective and intended use statements were deemed insufficient due to generality and vagueness. Evaluation objectives could be improved with extending understanding of intended use to describing who exactly may do what with specific evaluation information.

Evidence Sources: ALNAP Section 1 ratings; PrgES Feasibility, U3 and U35 ratings

Recommendation: Consider reducing the number of prescribed evaluation questions, or headlining these are more of a range of options, ordered by perceived importance, and the final list should be determined with guidance from external evaluators. Consider specifying primary intended user groups and users of evaluative information and what specific actions they may take (short of evaluation informed recommendations) if they received certain findings for certain evaluation criteria.

5. Methodology: Good

Definition: "Evaluation methodologies must be sufficiently rigorous such that the evaluation responds to the scope and objectives, is designed to answer evaluation questions and leads to a complete, fair and unbiased assessment."

SEU Description: The most frequent methods were document reviews (97%), key informant interviews (97%), focus group discussions (58%), secondary data analysis (52%), field visits (42%), and surveys (13%). Portfolio-wide accuracy and methods scores were 50% and 72% respectively. There were select instances where scoped methods were unable to sufficiently answer evaluation questions or speak to specific evaluation criteria, mostly in instances of investigating specific dimensions of outcomes and impact. But for the majority of evaluations, methods were appropriate to evaluation questions and purpose. Absent most evaluation inception reports and methodology sections were descriptions of evaluation design types or overarching evaluation approaches or models that the methods corresponded to. Data analysis descriptions were mixed, with some reports providing sufficient detail, and others no detail at all. Some managers mentioned instances of insufficient analysis methods.

Evidence Sources: PrgES Accuracy Ratings; ALNAP Methods Ratings; Transversal Methods Ratings

Recommendation: Continue the practice of requiring evaluation matrices that require evaluators to associate methods to data sources, indicators, standards, criteria, questions, and criteria. Ask more pointed questions about specific evaluation theories, approaches, or models that inform prospective evaluation team practices, and certainly for evaluation inception report methods sections. Consider revising the IR methods section accordingly to situate methods and procedures within approaches, models, and designs to ensure coherence of methods to evaluation purpose distinctions. Consider use of evidence rubrics in inception stage planning to see if evidence quality meets the needs of intended users and appropriate for the stakes of the evaluation.

6. Stakeholder engagement and reference groups: Very Good

Definition: "Inclusive and diverse stakeholder engagement in the planning, design, conduct and follow-up of evaluations is critical to ensure ownership, relevance, credibility and the use of evaluation. Reference groups and other stakeholder engagement mechanisms should be designed for this purpose."

SEU Description: the meta-evaluation team was pleased to learn all evaluation processes entailed the formation of a consultation group that assisted in the design, quality assurance, and responsiveness of the evaluation procedures. Evidence suggests the extent to which consultation groups are active in use and follow-up stages is low.

Evidence Sources: Participant Observation; Inception Reports; KIIs

Recommendation: continue utilizing consultation groups in all evaluation processes. Consider how to extend their involvement in the months following evaluation report delivery for use and follow-up activity improvement. Consider how to more fully include perspectives of groups external to MSF but internal to the evaluation context.

7. Human-rights based approach and gender mainstreaming strategy: Fair

Definition: "The evaluation design should include considerations of the extent to which the United Nations system's commitment to the human-rights based approach and gender mainstreaming strategy was incorporated in the design of the evaluation subject."

SEU Description: see Norm 8 description.

Evidence Sources: ALNAP Section 4.4

Recommendation: see Norm 8 recommendations.

8. Selection and composition of evaluation teams: Good

Definition: "The evaluation team should be selected through an open and transparent process, taking into account the required competencies, diversity in perspectives and accessibility to the local population. The core members of the team should be experienced evaluators."

SEU Description: Overall evaluator credibility ratings were "Good" across the portfolio. Managers expressed some challenges in rating the degree to which external consultants met competency expectations in the recruitment stage. Some reported delays in recruitment in finding the right match of consultants. There were multiple reports of less than ideal relations with evaluator or evaluation teams for some evaluation cases. One evaluation was canceled after the inception report quality assessment. Another case had no dissemination of the final evaluation report, which was deemed unsalvageable. Other cases involved considerable intervention from evaluation managers to improve quality of evaluation products. This is a challenging standard to strike a right balance in terms of selecting from among the available candidates the right evaluator or team for the evaluation. Further, multiple survey respondents and interviewees indicated the trade-offs of having someone with robust evaluation experience and little to no contextual experience and low evaluation skills and high understanding of the context.

Evidence Sources: Overall PrgES and ALNAP Ratings; Survey data; ToR; Participant observation

Recommendation: prioritize evaluation capacity and competency over contextual experience and subject-matter expertise. Identify internal subject-matter experts and referents to support and advise evaluation processes, and maintain robust inception stage procedures to enable external evaluators the time to orient themselves to the object of evaluation and its organizational, political, and cultural context. Refer to the ethics and human rights and gender equity norms and terms of reference standard recommendations for considerations of equity in recruitment. Consider augmenting the recruitment process to entail light-touch expressions of interest and conversations before detailed proposals are generated. Regardless of any modification to selection processes, make description of selection processes more transparent in TOR and require evaluation teams to discuss team composition and potential biases in inception and final reports.

9. Evaluation report and products: Good

Definition: "The final evaluation report should be logically structured and contain evidence-based findings, conclusions and recommendations. The products emanating from evaluations should be designed to the needs of its intended users."

SEU Description: portfolio-wide scores for report quality using the ALNAP standards were high with an average of 88% and a Very Good Rating. These are minimum standards, but indicative of overall report quality, resulting from templated reports, and multiple quality assessment points. PrgES Accuracy ratings are lower, with lower scores for the following sub-criteria: explicit evaluative reasoning, information management, reliable information, and justified conclusions and decisions. Also, of importance and related to the logic of evaluative reasoning, most all reports had criteria or dimensions of merit (though not all) but many did not have explicit setting of standards, or degrees of quality or performance, for each criterion. This resulted in various practices for reporting actual performance against these standards from binary met, not met, to meandering descriptions of mixed results with no conclusive judgment, to some using descriptors like good or very good, but without the comparison of a qualitative scale and equally applied across criteria within the case. A positive note about evaluation reports and products is the effort the SEU takes to create multiple report formats for meeting different audience needs, from slide decks, to full reports, to short versions, to posters, to reports in multiple languages.

Evidence Sources: SEU Report template; ALNAP Section 5; Overall PrgES and ALNAP Ratings; Qualitative notes from reviewers

Recommendation: Invite strengthening of evaluative reasoning by 1) underscoring existing guidance to connect findings, to questions and criteria; 2) link to evidence as much as possible such as in instances where findings were footnoted with evidence sources; 3) strengthen connection from findings to conclusions, where conclusions are not a discussion section of information not covered in findings; 4) ensure recommendations are tethered to specific conclusions or findings, and even potentially associated with criteria or evaluation questions. In short, overall strengthening connections of evaluative logic in reports can be improved, and likely met with ease with managers more fully articulating this value in templates and exchanges with consultants. A final and important recommendation we strongly suggest is to not stop at criteria in the General Logic of evaluation. Work with consultation groups and evaluators to establish standards of criteria, degrees of quality or goodness such as poor, fair and good, and even adequate levels of quality or goodness for each individual criterion, depending on the nature of the evaluation object and evaluation purpose. The most intuitive place to integrate this practice is in an additional column in evaluation matrices in inception report annexes. Finally, the practice of multiple reporting formats should be continued and continually improved to meet audience needs. A light-touch desk review of optional evaluation reporting formats and potentially the extent to which different formats are preferred and have or could meet different MSF audience needs could make this strategy more efficient and effective.

10. Recommendations: Good

Definition: “Recommendations should be firmly based on evidence and analysis, clear, results-oriented and realistic in terms of implementation.”

SEU Description: Recommendations were a feature of all but one evaluation with a total of 184 recommendations across 30 evaluations averaging 6 recommendations per evaluation. The PrgES does not have a standard for recommendations, likely resulting from a philosophical position about the necessity of recommendations in evaluations and more importance on evaluative conclusions. The ALNAP Proforma score for the recommendation standard was 94%, which is possibly inflated due to lack of deep contextual and organizational understanding from the report reviewers. Survey responses indicated recommendations are valued highly and likely the most scrutinized aspects of evaluations. Responses suggest recommendations in reports are often an unstated proxy indicator of quality including evaluator contextual understanding, which is also highly valued in OCB. Evaluators are equipped with adequate guidance in the evaluation report template and encouraged to make bold, but warranted recommendations in their reports and more recently a practice of collaborative recommendation making has been encouraged through evaluation policy and report template guidance. One evaluation report indicated any changes between recommendations of the initial and final drafts of the final report. No transversal analysis was conducted to determine the relative quality of recommendations by case, but many survey respondents across roles noted variation in applicability, specificity, feasibility, and other considerations. These mixed qualitative ratings are weighed with more importance than our external document review and this rating reflects that.

Evidence Sources: ALNAP Section 5.1.iii; Survey responses

Recommendation: Together with a special working group comprised of past consultation group members and intended evaluation users, the SEU might consider developing a recommendation matrix or rubric that formalizes the dimensions of recommendation quality and intended use, with such variables such as, but not limited to: ease of implementation; degree of anticipated benefit; program-specific or cross-cutting; feasibility; range of application; associated evaluation criterion, etc. Invite evaluators and consultation group members to assess and report assessments of recommendations according to these rubrics.

11. Communication and dissemination: Very Good

Definition: "Communication and dissemination are integral and essential parts of evaluations. Evaluation functions should have an effective strategy for communication and dissemination that is focused on enhancing evaluation use."

SEU Description: The SEU has a standard procedure of developing Use and Dissemination plans, with 71% (12/17) of evaluations having plans from the first instance of use in 2020. These are mostly communication and dissemination plans, but often do speak to use and utilization in terms of key meetings or decision points with specific activities to support operational management. The overall portfolio of evaluations for the 5 years saw 39% of evaluations with use and dissemination plans. Even in the absence of these plans, multiple reporting formats and reporting moments or events such as workshops, roundtables, and webinars have been a main feature. Where evaluations had specific Use and Dissemination Pans, many survey respondents indicated most of the activities were fulfilled. It is unclear the extent to which Use and Dissemination plans are used as management tools for follow-up.

Evidence Sources: Use and Dissemination Plans; PrgES P5 and A8 ratings; ALNAP Section 5

Recommendation: Continue to develop and use Use and Dissemination plans and recommit to using these plans as the map for significant re-engagement with increasing actual evaluation use and follow-up. We strongly encourage the SEU to reclaim the authority to follow-up with recommendations as well as specific activities that facilitate use.

5. QUALITY: GOOD

1. Quality assurance system: Good

Definition: "The head of evaluation should ensure that there is an appropriate quality assurance system."

SEU Description: The SEU recently joined a small club of evaluation units that have contracted and managed external independent portfolio-wide meta-evaluations (that are publicly shared). The commissioning of this meta-evaluation is an important indicator of commitment to quality assurance and was determined to systematically meet 6/18 evaluation accountability standards related to external meta-evaluation. The SEU does not have a practice of commissioning external meta-evaluations for discrete evaluations, and it is questionable how appropriate one-off external meta-evaluations would be given the average budget of individual evaluations. There are also clearly informal but regular quality checks for each evaluation at multiple stages of the evaluation. Evaluation manager interventions in analysis and reporting are evidence of these quality assurances functioning, as well as other ad hoc instances of managers intervening to strengthen processes and relationships in evaluations. Despite not matching standard definitions exactly, three of the six internal meta-evaluation standards were met in the spirit of the standard because of these clear quality assurance measures. Further, this meta-evaluation was greatly facilitated by decent evaluation documentation and coordination of artifacts that should be used for meta-evaluation. No explicit meta-evaluative framework exists but building blocks of quality evaluation practice is articulated across multiple SEU evaluation policy documents. Interestingly, the Annual Reports the SEU produces are a type of formal meta-evaluation, specifically a principles-focused evaluation model that looks at the coverage and fidelity of multiple evaluations of that year to effectiveness and operational principles articulated in the Strategic Orientations. This portfolio-level meta-evaluation could be replicated with the SEU using its own effectiveness principles of evaluation quality.

Evidence Sources: PrgES Evaluation Accountability Rating; KII; Management documents; Participant Observation

Recommendations: Formalize internal the meta-evaluation function by 1) recognizing and consolidating existing meta-evaluative criteria articulated in existing policy documents; 2) developing rubrics or score-cards for consultation groups to use to complement their qualitative reviews and comments for inception and final reports; 3) formalize internal meta-evaluation procedures to be guided by adapted checklists and rubrics from this portfolio-wide meta-evaluation. That is, identify evaluation case-specific criteria (not system-level dimensions) and standards that are important to the SEU/OCB and develop fit-for-purpose and fit-for-scope review instruments for managers; 4) consider developing stage-specific checklists with respective quality standards from the 180 PrgES standards that were coded by SEU evaluation stage. Quite possibly, one of the most important meta-evaluative standards we could strongly encourage is the systematic planning for and collection of data about the consequences of evaluations including stakeholders use of findings. This is a moment to be the change many of your evaluations reported having limitations with, that is adequate monitoring data about intervention effects. The SEU failing to investigate the effects of evaluation interventions is comparable to OCB failing to investigate the effects of medical humanitarian interventions. This incongruence should be rectified through concrete plans for remediation.

2. Quality control of the evaluation design: Good

Definition: “Quality should be controlled during the design stage of evaluation.”

SEU Description: Setting aside most of the evaluation inception reports do not actually designate a specific evaluation design type, but instead a collection of methods, most inception reports are sufficiently detailed, with minor variation. Some evaluators expressed dissatisfaction with the lengthy inception stage encouraged by the SEU, but our position is that these preliminary activities are valuable and contributive to evaluation success.

Evidence Sources: PrgES Evaluation Accountability Rating; KII; Management documents; Participant Observation

Recommendation: Headline to evaluation constants typical evaluations at SEU have in-depth inception stages for contextual orientation and quality assurance. Create a new section, or add further guidance in inception reports that expects clearly articulated evaluation design, or any relevant evaluation theory, approaches, or models that inform and guide the selection of methods and activities for supporting intended use. Special attention can be given to reducing variation of interpretation of information requests in evaluation matrices (distinguishing between how evaluators interpret columns of the matrix such as judgment criteria, indicators, evaluation domains). Consider developing a rubric for consultation groups to use in addition to ad hoc commentary on inception reports.

3. Quality control at the final stage of evaluation: Fair

Definition: “Quality should be controlled during the final stage of evaluation.”

SEU Description: in-built feedback loops for the final stage of evaluation are the consultation group and manager comments on initial final drafts, a working session with the consultation group, and revisions to the final report. Additional reporting formats are developed by managers and the SEU coordinator to meet different audience needs. There is however a clear lack of management responses as well as recommendation follow-up, which the meta-evaluation team views as key factors in attesting to report quality, if not quality control in the case of management responses, and ensuring use quality for recommendation follow-up.

Evidence Sources: PrgES Evaluation Accountability Rating; KII; Management documents; Participant Observation

Recommendation: We strongly encourage a recommitment to meaningful management responses that are not seen as mere formalities, but as functional documents. We strongly encourage more resources dedicated from the SEU to follow up with evaluation use and consequences starting with a recommitment to recommendation follow up procedures, as well as evaluation consequence inquiries possibly at 3, 6, and 12 months for each evaluation, depending on the nature and scope of the evaluation. Consider developing a user-friendly final report rubric for consultation group members to use in their reviews as a more systematic form of internal peer review. Consider having the management response entail a final attestation to report/process quality in a rubric that is shared in the final report annex.

>-<

ANNEX VII: PORTFOLIO AND SEU SYSTEM ASSESSMENT OF SEU-SPECIFIC ELEMENTS OF EVALUATION QUALITY: GOOD

This document contains meta-evaluative judgments about the portfolio of evaluations from 2017-2021 and SEU evaluation system using the various professional values, quality domains, criteria, standards, and effectiveness principles as articulated across multiple SEU guideline and policy documents.[78] Each quality element contains a title, rating, a brief description of actual performance, evidence sources, and, if applicable, element-specific recommendations. This quality framework was not applied to each individual evaluation, but used ratings at the aggregated portfolio-level along with other sources of evidence to reach these evaluative conclusions about the SEU system and past 5 years of evaluation performance. The levels of performance quality are the same levels used in the PrgES and ALNAP Proforma frameworks (Excellent, Very Good, Good, Fair, and Poor). These ratings are not based on a predetermined ratio of met or not met standards and a systematic formula for synthesis, but on a generic qualitative rubric shared across all elements as shared below and interpreted by the meta-evaluation team.

QUALITY ELEMENT RUBRIC

Excellent	**Very Good**	**Good**	**Fair**	**Poor**
Element is completely manifest in actual performance and supported by strong evidence.	Element is mostly manifest in actual performance and supported by strong evidence.	Element is partially manifest in actual performance as indicated by sufficient evidence.	Element is partially to mosty not manifest in actual performance as indicated by sufficient evidence.	Element is mostly or completely not manifest in actual performance supported by strong evidence, or there is no evidence to support claims of meeting this quality element and this lack of evidence is treated as lack of performance quality element in action.

MSF, OCB, AND SEU VALUES OF EVALUATION

As a movement, MSF values the evaluation function, or disciplined systematic value-based inquiry, as evidenced by the high-level policy document The La Mancha Agreement. The Agreement states that MSF makes a “commitment to evaluation” and “aspires to ensure quality and relevance in operations, is committed to the impact and effectiveness of its work so that good work can be multiplied and abandon ineffective practice.”[79] The Agreement also acknowledges MSF values accountability and

[78] These include: SEU Evaluation Framework; SEU Evaluation Manifesto; The Six Step Process for Evaluations; SEU Ethical Guidelines; SEU Roles and Responsibilities; SEU OCB Strategy and Governance; SEU Steering Committee Terms of Reference.

[79] SEU Evaluation Manifesto.

transparency "to those we assist, our donors and wider public." OCB has translated this high-level commitment to accountability and transparency with a commitment in its 2020-2023 Strategic Orientations[80] to a "culture of evaluation" that "give[s] the field teams the opportunity to learn from [their] practices and to constantly improve the quality and pertinence of operational/medical interventions." OCB has subsequently adopted an evaluation policy that all projects should be evaluated during their lifespan unless there is a well justified reason not to.

Through inception stage key informant interviews we did establish that definitions of quality and emphases placed on different dimensions of quality differs across the organization and between individuals, including those within the SEU. However, claims about evaluation quality from values, principles, and standards are dispersed across multiple policy and guidance documents, which in concert constitutes an emerging quality framework, albeit inchoate. These are the professional values related to or about evaluation that are espoused across multiple policy documents that have been categorized by the SEU's conceptual domains of quality:

Value	**Use**	**Methods**
Transparency Accountability Downward Accountability Credibility Objectivity Independence Necessity Impartiality Ethics Honesty and Integrity Inclusivity Engagement and Ownership Respect for Dignity and Diversity Avoidance of Harm	Utility Use Effective Evaluation Processes Learning Real Time Learning Follow up on Findings and Recommendations Culture of Evaluation	Rigor Accuracy Completeness Reliability Confidentiality Quality Assurance/Control

All of the stated values here are directly duplicated or associated with criteria, quality domains, and norms and standards in the PrgES, ALNAP, or UNEG Frameworks. Ratings of the SEU evaluation portfolio for the past 5 years and the SEU evaluation system for each of these values will not be given here to avoid duplication. However, ratings for the emerging quality framework communicated in the Evaluation Manifesto Guideline document will be provided below.

The meta-evaluation terms of reference included the following disclaimer: "There is currently no formal adopted framework of quality in the evaluations managed by the SEU on behalf of OCB although the work of the unit is influenced by several frameworks including the Joint Committee on Standards for Educational Evaluation (JCSEE) Program Evaluation Standards, ALNAP Proforma, as well

[80] Interestingly, the new SEU Annual Reports function as Principles-focused Evaluations of Strategic Orientations at the OCB. This is a form of formalized meta-evaluation that explores the extent to which effectiveness and operational principles within the Strategic Orientation were manifest in evaluations.

as various evaluator competency frameworks, including those from the American Evaluation Association and the United Nations Evaluation Group (UNEG). It is likely that ideas on what constitutes quality or value for different stakeholders in evaluation within the context of MSF and OCB differ across the organization. It will be necessary to establish a framework of accepted criteria as part of the evaluation process." This segment is drawn from the Evaluation Manifesto and prefaces a description of a collection of aspects or elements of quality that the SEU understands as constituting quality in evaluation, grouped into three domains: value, use, and methods, which can certainly be considered a partial, yet formal framework for evaluation quality.

SEU EVALUATION MANIFESTO QUALITY FRAMEWORK RATING: GOOD

Again, as explained in the manifesto, this framework is informed by the Program Evaluation Standards, the ALNAP Proforma, AEA evaluator competencies, and UNEG Norms and Standards. It is composed of various elements that register what quality means to the SEU, which are expressed as domains (broad categories), activities (evaluation processes), principles (prescriptive actions), products (documents), or events. For sake of brevity, the definitions for each of these elements will not be shared as in the UNEG annex, but can be viewed in detail in the SEU Evaluation Manifesto. Domain ratings are based on average sub-ratings that are rounded to the nearest whole number and rounded up at half points. Ratings of elements and sub-elements under domains were averaged with the following numerical codes (poor=1; fair=2; good=3; very good=4; excellent=5).

VALUE (DOMAIN): GOOD

Choosing Criteria (activity): Very Good

SEU Description: Across the portfolio, 97% (n=30/31) of evaluations used at least one OECD-DAC criterion, with the exception of an evaluation that investigated an OCB budget overspend. The average evaluation used a combination of 3 OECD-DAC criteria. The overwhelming majority of evaluations the SEU manages are goal-oriented, with 87% (n=27/31) investigating effectiveness. Other regularly occurring OECD-DAC criteria were Relevance (68%, n=21/31), Efficiency (58%, n=18/31), Impact (52%, n=16/31), Sustainability (16%, n=5/31), and Coherence (6%, n=2/31). Among all evaluations, 94% (n=29/31) used additional evaluation criteria, with the most common criterion being "Appropriateness" occurring (51%, n=16/31) of the time. These findings suggest an appropriate balance of generally accepted criteria, and responsiveness to specific evaluation needs and contexts with custom criteria. There was evidence to suggest in the ALNAP 4.3 standard that some criteria were not evaluable based on a number of factors, but mostly data availability.

Evidence Sources: evaluation portfolio review.

Ask the right questions (principle): Good[81]

SEU Description: The SEU averages prescribing about 15 evaluation questions and sub-questions in their terms of reference, or 5 main evaluation questions and 3 sub-questions. Cursory analysis suggests evaluation questions are matched with criteria, and deliberated and adapted by evaluators on a case-by-case basis. Evidence also suggests some questions may not be answerable due to evaluability issues, especially those pertaining to impact or outcomes in some instances with limited data and effectiveness in instances without program logic and clear objectives. It is likely the SEU may be ambitious in the number of evaluation questions that are prescribed for consideration and more balance to depth and breadth could be considered.
Evidence Sources: final evaluation reports.

Engagement and ownership (domain): Good

SEU Description: the portfolio of evaluations scores highest in Utility among all other criteria in the PrgES, which houses many sub-criteria about evaluation participant engagement. Propriety is another criterion that pertains to this, which received a "Good" rating. Given this element includes ownership, there are reports of less than ideal procedures around downward accountability with all right-to-know audiences. Further, there is evidence to suggest that evaluation users have low degrees of owning intended use plans.
Evidence Sources: PrgES Utility and Propriety ratings; surveys.

Engage the voices of those less present (principle): Poor

SEU Description: the second checkpoint of sub-criterion U2: Attention to stakeholders states: "Search out and invite input from groups or communities whose perspectives are typically excluded, especially stakeholders who might be hindered by the evaluation." This indicator or checkpoint had only two evaluations attempt to do this as indicated in the IR in one and attempted but failed in the FR of another.
Evidence Sources: PrgES Utility sub-criterion 2 and 4.

Languages (domain): Good

SEU Description: at least one evaluation report was written in both French and English. One evaluator provided feedback in a survey that having a French speaking manager in the SEU would be ideal. We did not check to see if there are no French speakers, or just no native-French speakers. There is not conclusive evidence that the English reports could not or should not have been translated into other languages for in-country communication and use.
Evidence Sources: final evaluation reports

Ethics(domain): Good

SEU Description: see UNEG Norm 6 for a detailed description. It should be noted that PrgES P3 "Adhere to applicable federal, state, local, and tribal regulations and requirements, including those of Institutional Review Boards, local/tribal constituencies, and ethics committees that authorize consent

[81] A more detailed analysis about evaluation question quality was discussed at some point during the meta-evaluation as a transversal line of inquiry in service of meta-evaluation question 4 about factors of quality, but was ultimately not pursued due to other analytical priorities. This may be a worthwhile internal meta-evaluative transversal analysis in service of improving the design, feasibility, and utility of future evaluations.

for conduct of research and evaluation studies" was still rated and received a "Fair" rating despite an MSF Research Ethical Framework policy that has been interpreted to an evaluation policy that exempts SEU evaluations from ethical reviews.
Evidence Sources: UNEG Norm 6 evidence sources; PrgES P3.

USE (DOMAIN): GOOD

Learning (domain): Fair

Real time learning (domain): **Good**

SEU Description: the definition for this domain actually comes close to describing process use, or the benefit accruing to individuals due to their participation in evaluations. There is evidence from survey responses and interview transcripts that this type of learning has and does regularly occur. What is less apparent is if real-time learning extends to instrumental findings use, where formative evaluations share findings that assist operations with adaptive management.
Evidence Sources: survey data; key informant interviews.

Follow up on findings and recommendations (principle): **Poor**

SEU Description: See UNEG Norm 14. There were recommendation follow up documents for 3/31 evaluations and management responses for 7/31 evaluations. Evidence from project contacts and commissioners was mixed in terms of integration of findings and recommendations into planned action points. There was no existing documentation of evaluation consequences.
Evidence Sources: UNEG Norm 14 sources.

Link to strategic platforms and meetings (principle): **Good**

SEU Description: interview, survey data, and annual reporting do suggest findings from many evaluations are discussed at these key junctures, and in some instances for years after the fact. This is an example of evaluation influence which could be explored further.
Evidence Sources: KIIs; surveys; SEU annual reporting.

Communicate and disseminate findings (principle): Good

Cross project and Inter-OC learning (domain): **Good**

SEU Description: activities related to this domain such as sharing on Interna Ops Newsletter and Inside OCB are found across many use and dissemination plans, with surveys reporting these and other related activities such as webinars. Reports from the 2021 however suggest lack of activity in the Intersectional Evaluation Group, which presumably corresponds to this domain.
Evidence Sources: UandD Plans; survey; interviews; SEU reporting.

External communication (domain): **Very Good**

SEU Description: all evaluations are assumed to be shared and publicly available per activities in UandD plans. Some evaluations were checked for public accessibility by reviewers, but not all.
Evidence Sources: UandD plans; evaluation contracts; evaluation policy documents.

Transversal Learning (domain): Excellent

Annual report (product): **Excellent**

SEU Description: evidence[82] from these reports demonstrate they exhibit "Excellent" potential for *Transversal Learning.* These reports that investigate the coverage of operational priorities and fidelity to strategic orientations are exceptional internal evaluative reports. This is one of the few instances of the SEU actually conducting their own evaluations, as opposed to managing them, and these limited examples are of high quality.

Recommendation: consider including a section or emphasis that makes similar meta-evaluative judgments about that year's evaluations a revised EMQF based on effectiveness principles. Consider integrating aspects of Principles-focused Evaluation to support the conduct and delivery of these unique internal evaluation reports with regards to existing domains of operational priorities and strategic orientations.

Evidence Sources: annual reports.

Evaluation day (event): **Excellent**

SEU Description: the evaluators were able to review some artifacts from evaluation days, including transversal analyses. This annual event signals an enabling evaluation environment, a commitment to promoting evaluation culture. From the outside looking in, evidence from these events suggest an excellent effort and performance for stimulating transversal learning.

Evidence Sources: transversal analyses; SEU reporting; interviews.

Annual presentation and discussion at OCB board (event): **Very Good**

SEU Description: discussions with the head of unit suggest these presentations can be limited in attempting to distill the activities and outputs of the full dossier into a few bullet points on a slide deck. However, this rating is mostly based off the honest and transparent reports found in the SEU reporting documents, not the stand alone principles-focused Annual Reports that started in 2020, but synthesis of quarterly reports. Review of these reports show honest disclosure about challenges and successes and show an upward trajectory in terms of demand, credibility, and quality of evaluations through the years.

Evidence Sources: SEU reporting documents.

METHOD (DOMAIN): GOOD

DATA (DOMAIN): GOOD

SEU Description: a known and recurring issue within OCB is the lack of consistent monitoring data at the project level including the lack of clear program design theory that would dictate appropriate and needed types of data for management, let alone evaluation. However, evaluations regularly make use of existing data with 53% of all evaluations using secondary data analysis as a data collection and analysis method. Despite limitations, many evaluations used these existing routine health data to arrive at defensible and laudable data-informed evaluative judgments about effectiveness. Primary data collection methods were predominantly qualitative in nature with interviews, focus groups, and document reviews serving as primary data sources.

[82] Annual reports.

Evidence Sources: inception and final reports.

CONSIDER THE EVALUABILITY OF THE PROJECT (PRINCIPLE): POOR

SEU Description: See UNEG standard 4.2.

Evidence Sources: UNEG standard 4.2 sources.

DISCUSS EVALUATOR COMPETENCIES (PRINCIPLE): VERY GOOD

SEU Description: see UNEG standard 3 description.

Evidence Sources: UNEG standard 3 sources.

>-<

ANNEX IIX: CREDIBILITY RATINGS OF FRAMEWORKS, EVIDENCE, AND ANALYSES

The following is an internal assessment of the credibility of the meta-evaluation frameworks, analytical procedures, and evidence sources applied and used in this study. The assessment used an adapted *Quality of Evidence Rubric* by Thomas Aston. The following accuracy criteria from the rubric were adapted—triangulation, transparency, and independence—and were supplemented by a fourth criterion of reliability. These rubrics and judgements are reported to assist report readers in contextualizing and qualifying meta-evaluation claims. Readers could use the same rubric and arrive to different ratings based on differences in ratings or use different or additional criteria to arrive at difference judgements than these. The decision to use the same four criteria for credibility across different sources of credibility issues, instruments, data sources, and procedures means the relevance of criteria was not always the same for each object being judged.

Criteria	1	2	3	4	5
Triangulation	No evidence corroborates the connection between standard and judgement. Other data contradict the proposed connection.	A single source of evidence makes the claim.	Multiple lines of evidence corroborate the connection between standard and judgement.	Multiple lines of high-quality evidence corroborate the connection between standard and judgement.	Multiple lines of evidence across mixed methods corroborate the connection between standard and judgement.
Transparency	Explicitness of framework and analytical procedure not described clearly. It is unclear what evidence supports the claim.	Explicitness of framework and analytical procedure not description low. Evidence has been identified, but not clearly explained.	Explicitness of framework and analytical procedure not description moderate. Various sources of evidence are clearly identified and explained.	Explicitness of framework and analytical procedure not description high. Sources of evidence and data collection methods are clearly explained. Data limitations and alternative interpretations are clearly discussed.	Explicitness of framework and analytical procedure not description very high. Sources of evidence and data collection methods are clearly explained. Data limitations and alternative interpretations and the plausibility of alternative explanations are clearly discussed. Data collection protocols and raw data is available.
Independence	Evidence is self- reported. Sources are closely connected to the intervention and known to have significant biases and strong incentives to	Evidence is self-reported. Primary and/or secondary data indicate a potential lack of independence and number of potential biases.	Evidence may be collected by partners or collected by independent evaluators. Issues of potential bias are unknown.	Evidence is collected by independent evaluators without clear connections to the intervention. Sources of potential bias are clearly signposted, and efforts have been made to limit these.	Evidence is collected by independent evaluators without clear connections to the intervention. Sources of potential bias are clearly signposted and considerable efforts have been made to limit these. The account is corroborated

	potentially misrepresent the events.				by those closest to events but with no known connection to the intervention.
Reliability	No replicability of framework or procedures due to lack of transparent elements or steps. Evidence source credibility non-existent.	Replicability of framework or procedure low due to nature of descriptions. Evidence source credibility and trustworthiness low.	Replicability of framework or procedure moderate due to nature of descriptions. Evidence source credibility and trustworthiness moderate.	Replicability of framework or procedure high due to nature of descriptions. Evidence source credibility and trust-worthiness high.	Very High degree of framework application or analytical procedure replicability; evidence sources are highly credible due to proximity, positionality, and trustworthiness.

Quality Frameworks

PrgES

Criteria	Rating	Explanation
Triangulation	5	Each evaluation rated by this framework derived it's overall rating from multiple evaluation artifacts, as well as a survey sent to those most involved with the evaluation. In addition, two calibrated reviewers assessed the portfolio, while a third did quality checks of the other two's work.
Transparency	5	Sources of data collection, and the process of using this framework are clearly explained (see Annex II). Limitations are discussed, and raw data is provided. Justifications and evidence sources for judgements are provided.
Independence	5	Both the framework itself, and the evaluators rating based on the framework, are separate from the SEU and MSF.
Reliability	5	Detailed indicators, instructions for checklist application, use of cut scores, and formulas for numerical weight and sum synthesis for each criterion translated to high replicability. Prior research has demonstrated issues with interpretation between raters and issues with process indicators. These known issues were addressed with significant calibration stage between raters and development of detailed codebook and formative and summative reliability checks and assessments.

ALNAP

Criteria	Rating	Explanation
Triangulation	4	Each evaluation rated by this framework derived it's overall rating from multiple evaluation artifacts. In addition, two calibrated reviewers assessed the portfolio, while a third did quality checks of the other two's work. Surveys did not factor in this framework as a working assumption of the associated checklist is that all standards would be evident in reports. Due to only a quarter of the indicators used in the PrgES and only the use of artifacts, this scores one lower than PrgES for this criterion.
Transparency	5	Sources of data collection, and the process of using this framework are clearly explained (see Annex II). Limitations are discussed, and raw data is provided. Transparency of synthesis would have been lower if we used the prescribed grading system as opposed to ratings based on scores and cut scores. Justifications and evidence sources for judgements are provided.
Independence	5	Both the framework itself, and the evaluators rating based on the framework, are separate from the SEU and MSF .
Reliability	4	ALNAP recommends the use of dual raters and synthesis of grades but doesn't provide explicit rubric for what constitutes an A versus an F grade for a quality dimension, neither inter-rater reliability measures, nor ideas for resolving conflicts. We addressed these limitations by using the PrgES method of numerical weight and sum with cut scores. There is less detail in dimensions than PrgES, which suggests reliability could be lower than PrgES.

UNEG

Criteria	Rating	Explanation
Triangulation	5	Data sources for this framework came from both the PrgES and ALNAP checklists, giving substantial triangulation as well as survey data, interviews, and participant observation.
Transparency	4	Justifications and evidence sources for judgements are provided, but ratings were not derived from the presence or absence of qualitative indicators whose ratios were summed and synthesized to scores, and ratings applied to cases with use of cut scores, where cases were then averaged for portfolio scores and ratings. Further no rubric for each Norm and Standard were not used, but descriptions of those Norms and Standards are clear.
Independence	5	Both the framework itself, and the evaluators rating based on the framework, are separate from the SEU and MSF.

Reliability	3	Only one reviewer made initial judgments, while two additional team members assessed and confirmed the conclusions made. Similar constructs to the PrgES and ALNAP were compared with those ratings. Additional dimensions related to the evaluation system were informed by observation and additional data collection. A generic rubric of standards was used, but not explicated in detail on what distinguishes one degree from the next, not for each norm or standard. There is a higher chance for differences of interpretation given ratings were derived from a qualitative synthesis of data and not an algorithm of weighting and summing scores and use of cut scores.

EMQF

Criteria	Rating	Explanation
Triangulation	5	Data sources for this framework came from both the PrgES and ALNAP checklists, giving substantial triangulation as well as survey data, interviews, and participant observation.
Transparency	4	Justifications and evidence sources for judgements are provided, but ratings were not derived from the presence or absence of qualitative indicators whose ratios were summed and synthesized to scores, and ratings applied to cases with use of cut scores, where cases were then averaged for portfolio scores and ratings. Further no rubric for each Norm and Standard were not used, but descriptions of those Norms and Standards are clear.
Independence	3	While the team member reviewing the data was independent, the framework itself comes from the SEU. This does increase the relevance of the framework and does not undermine justifications for application. Further, the framework was not exactly viewed as a framework by the SEU, but a looser collection of ideas about evaluation not formalized and likely not intended for measurement use, which might help make the case for more independence in its application.
Reliability	3	Only one reviewer made initial judgments, while two additional team members assessed and confirmed the conclusions made. Similar constructs to the PrgES and ALNAP were compared with those ratings. Additional dimensions related to the evaluation system were informed by observation and additional data collection. A generic rubric of standards was used, but not explicated in detail on what distinguishes one degree from the next, not for each norm or standard. There is a higher chance for differences of interpretation given ratings were derived from a qualitative synthesis of data and not an algorithm of weighting and summing scores and use of cut scores.

Analysis Methods

Numerical Weight & Sum Methodology

Criteria	Rating	Explanation
Triangulation	5	The NWS synthesis method for case-level analysis (PrgES and ALNAP) and portfolio-level analysis (UNNEG and EMQF) used multiple sources of evidence and multiple types of data.
Transparency	4	NWS application for frameworks that had indicators and scoring was highly transparent, and less so with synthesis of dimensions and sub-dimensions where only ratings were provided. Qualitative ordinal data on degrees of performance were coded numerically and averaged to facilitate synthesis, but the combination of facts and values for rating-only portfolio judgements is less transparent than those of cases, despite descriptions and evidence source disclosure.
Independence	5	Procedure applied by external evaluators.
Reliability	4	Among limitations with methodology, assumption of score commensurability likely most relevant to reliability. Resolution to this issue and others is to use qualitative weight and sum methodology but was opted against for ease of interpretation and feasibility with portfolio scope.

Cost-Utility Analysis

Criteria	Rating	Explanation
Triangulation	4	Utility data came from artifacts and surveys. Budget data came from external consultant budgets. Lower rating reflects lack of integration of institutional budget data from SEU overhead and human resources.
Transparency	4	Clear explanations of methods are provided, but little discussion on potential limitations or alternatives definitions on this specific method.
Independence	4	All analysis was completed by independent evaluators, but dimensions of potential bias on this discrete method were not explored.

Reliability	4	Utility measures biggest potential source of measurement error or unreliability, though risk is low.

Historical Analysis

Criteria	Rating	Explanation
Triangulation	5	Quality scores derived from PrgES, which is most triangulated framework.
Transparency	4	Score derivation and evidence sources clearly described. Specific measure of association for correlational statistic not reported.
Independence	5	Procedure applied by external evaluators.
Reliability	5	Procedure explained in detailed methods note.

Quality Gap Analysis

Criteria	Rating	Explanation
Triangulation	5	Derived from PrgES scores, the highest triangulated scores.
Transparency	5	Clear description of how scores were derived is provided.
Independence	5	Procedure applied by external evaluators.
Reliability	4	Interpretation differences could occur at the indicator level between different coders. This study’s inter-rater reliability scores were within acceptable bounds.

Max Deviation Analysis

Criteria	Rating	Explanation
Triangulation	4	Explanations of contributing factors were coded solely from open-ended survey responses, across 10 of the 32 evaluations. Multiple survey responses were used for each one if available.
Transparency	3	Procedures explained, but no codebook provided.
Independence	3	All analysis was completed by independent evaluators, but self-report data by those with close proximity to the evaluation cases poses risks in recall bias and social desirability bias.
Reliability	3	No codebook results in limited replicability and consistency.

Limitations Analysis

Criteria	Rating	Explanation
Triangulation	3	Limitations we're coded across all 31 evaluation artifacts. Limitations limited to those identified by external consultants.
Transparency	3	Procedures explained, but no codebook provided.
Independence	3	All analysis was completed by independent evaluators. Self-report of limitations by those who may have had responsibility in identifying and mitigating limitations beforehand could be subject to social desirability bias.
Reliability	3	No codebook results in limited replicability and consistency.

Evaluation Role and Responsibilities Analysis

Criteria	Rating	Explanation
Triangulation	4	Triangulation occurred across four roles, though high variation in distribution of responses by role.

Transparency	3	Procedures explained, but no codebook provided.
Independence	3	Analysis conducted by evaluators. Surveys were confidential, but not anonymous, which could have led to social desirability biases.
Reliability	3	No codebook results in limited replicability and consistency.

Use and Use Outcome Analysis

Criteria	Rating	Explanation
Triangulation	3	Triangulation occurred across four roles and multipe survey items, though high variation in distribution of responses by role.
Transparency	3	Procedures explained, but no codebook provided.
Independence	3	Analysis conducted by evaluators. Surveys were confidential, but not anonymous, which could have led to social desirability biases.
Reliability	3	No codebook results in limited replicability and consistency.

Evidence Sources

KII's/FGD's

Criteria	Rating	Explanation
Triangulation	3	Decent variety of respondents, mostly from headquarters, with limited access to project-level staff and no patient or community input.
Transparency	4	Roles shared, but not the identities of respondents.

Independence	4	All interviews we're conducted by independent evaluators, and potential limitations or bias are clearly described in this report.
Reliability	4	Survey protocol is clearly described in Annex IX.

Evaluation Artifacts

Criteria	Rating	Explanation
Triangulation	5	A wide variety of documents were used across numerous evaluations, each of which we're the product of collaboration amongst numerous stakeholders and evaluators
Transparency	4	Included evaluation artifacts are clearly described in Methods and Annex II, but rationale for why some were or were not included was not clear.
Independence	2	All artifacts are collected as secondary data managed by those who were part of evaluation object. Evidence of some artifact types not shared and no explanation for why these were not.
Reliability	5	No variation in

Survey Responses

Criteria	Rating	Explanation
Triangulation	3	While surveys were sent out to multiple respondents, several evaluations received little to no responses. Survey response rates reported elsewhere.
Transparency	4	Sampling frame described in number of surveys sent by role along with responses. Identities of respondents not shared.
Independence	4	Survey was conducted by independent evaluators, even though those responding we're those close to the evaluation. Potential limitations we're engaged and highlighted in limitations section.

Reliability	3	Survey instrument was not validated but revised after field piloting and feedback from initial respondents. Some evidence suggested issues with question interpretation.

Observations

Criteria	Rating	Explanation
Triangulation	3	Each evaluator served as an “evidence line” using observations in team meetings to shape findings.
Transparency	2	Observations are included in various findings but annotated and stored across various digital and non-digital locations, if at all as the totality of experiences by the meta-evlauation team were not all documented, but discussed and shared in collective memory.
Independence	3	All evaluators are independent of MSF, but also affected by the degree of quality of our own conduct of the meta-evaluation.
Reliability	2	Case experience of meta-evaluation unique, no prompts or observation protocol were developed and degree of observation from participation was informal.

ANNEX IX: SEMI-STRUCTURED KEY INFORMANT INTERVIEW PROTOCOL

30 to 45-Minute Protocol for Consultation Group members. Consider sharing the following itinerary with the interviewee in the call chat:

1. General Introduction (~5 minutes)
2. Professional Background (~5 minutes)
3. Views about Evaluation (~10 minutes)
4. Evaluation within the OCB (~10 minutes)
5. Criteria of Quality Evaluation/Research (~10 minutes)
6. Use of this Meta-evaluation (~5 minutes)

GENERAL INTRODUCTION

- We have been commissioned by the SEU steering committee to conduct a meta-evaluation, an evaluation of evaluations that SEU has managed for OCB for roughly the last 5 years. We are specifically looking at the quality of evaluations and the value of those evaluations for OCB. While we are within an inception phase of this work, we are identifying and engaging with various groups and individuals at OCB who have been, are, or could be users of evaluations and prospective users of this meta-evaluation. Our purpose in connecting is to help facilitate an agreement on the criteria and standards by which we will evaluate past evaluation performances and products. We are interested in hearing more about your professional values, views on evaluation, experience with evaluation with the SEU/OCB, ideas about evaluation quality, and potential uses of this meta-evaluation.
- From our interviews and written engagement, we are not disclosing the identity of respondents within the inception report that we will deliver to the consultation group and meta-evaluation users. We may quote text segments in supporting claims, but will not attribute those quotes to respondents and interviewees.

PROFESSIONAL BACKGROUND

1. From your experiences that led you to MSF and from your time at MSF, what would you say are the professional values (and by this, I mean beliefs about what is important) that you try to live by and make manifest in your work?

VIEWS ABOUT EVALUATION

2. When you hear or see the word evaluation, what comes to mind?
3. What is your relationship to evaluation? Do you get involved in them at MSF? Commission them? Use them in your work? Something else?
4. Can you think of a time at MSF when you had a positive experience with an evaluation, either as an evaluation participant or an evaluation user? What was that and what made it positive or successful in your eyes?

5. Conversely, can you think about a time at MSF when an evaluation did not go well? What made it so?

EVALUATION WITHIN THE OCB

6. Why do you think OCB commissions evaluations?
7. Do you have a sense of how evaluations have been used in the past by OCB? Have you used evaluations in the past? If so, what did that look like?
8. Is there something you wish OCB did more of in terms of commissioning, conducting, or using evaluations?

CRITERIA OF QUALITY EVALUATION

9. How would you know if an evaluation process and or product was successful? What would it look, sound, and or feel like?
10. What criteria or dimensions of quality or merit come to mind when you think about exemplary evaluation work?

USE OF META-EVALUATION

12. What do you expect this meta-evaluation process to reveal?
13. What do you not know now about the quality and value of evaluations at OCB that if you knew would be useful for you in your current role?
14. How do you hope this meta-evaluation is used within the SEU and OCB?

>-<

ANNEX X: META-EVALUATION SURVEY QUESTIONNAIRE

Dear Respondent,

As part of an ongoing meta-evaluation, you are being invited to participate in this survey because of your involvement in the **[evaluation code]** evaluation that the Stockholm Evaluation Unit managed for the Operational Centre Brussels of Médecins Sans Frontières. Responses are not anonymous. There are no benefits and no known risks for participating in this survey. For evaluators, evaluation commissioners, and project contacts, the survey should take 10-15 minutes to complete. For evaluation managers the survey should take around 30 minutes to complete.

Any questions about the survey can be sent to Tian Ford at tian@pointedarrows.com.

Thank you,

Pointed Arrows Consulting

1. Email:

2. I consent to take part in this meta-evaluation survey.
   a. Yes
   b. No
3. What was your role with the evaluation in the title of this survey?
   a. Evaluator
   b. Evaluation Manager
   c. Evaluation Commissioner
   d. Project Contact

PROJECT CONTACT BLOCK

Our records indicate that you were the project contact for this evaluation. The following questions ask you about your experience with the evaluation and for your feedback.

Any questions about the survey can be sent to Tian Ford at tian@pointedarrows.com.

1. How satisfied were you with this evaluation process?
   a. Very dissatisfied
   b. Dissatisfied
   c. Somewhat dissatisfied
   d. Somewhat satisfied
   e. Satisfied
   f. Very satisfied

2. Please explain your satisfaction level.
    a. (open-ended)
3. Which MSF-specific principles seemed evident in the conduct of this evaluation? (Example documents that contain specific principles include the *MSF behavioral commitments, SEU ethical guidelines, the Charter, the Chantilly Principles, or La Mancha Agreement)*
    a. (open-ended)
4. How was the evaluation disseminated and used?
    a. (open-ended)
5. Which of the following activities from the evaluation Use and Dissemination Plan took place, that you are aware of? (check all that apply)
    a. Recommendations follow-up
    b. ARO discussions and preparation, round tables, project design
    c. Share the report with DRC and other missions/cells
    d. Evaluation Poster + Report (prepared by SEU) – mail distribution
    e. Short update @ Flash Info/Info Matin
    f. Upload evaluation report on evaluations.msf.org (public)
    g. Presentation to the mission
    h. Presentation and Discussion Session Webinar within OCB
    i. Sharing the report with MOH and partners
    j. Presentation/Discussion with external stakeholders in DRC
6. How satisfied were you with this evaluation's use and dissemination?
    a. Very dissatisfied
    b. Dissatisfied
    c. Somewhat dissatisfied
    d. Somewhat satisfied
    e. Satisfied
    f. Very satisfied
7. Please explain your satisfaction level.
    a. (open-ended)
8. If this evaluation was used, what were the 1 to 3 most valuable outcomes of that use? Write NA if no use.
    a. (open-ended)
9. What were the most important factors that determined the outcome of this evaluation?
    a. (open-ended)
10. How are project contacts responsible for the success or failure of evaluations at OCB?
    a. (open-ended)
11. Please share any additional thoughts about the quality or value of this evaluation's processes or products not captured in the previous questions.
    a. (open-ended)

EVALUATION COMMISSIONER BLOCK

Dear Evaluation Commissioner,

Thank you for agreeing to participate in this survey. Your insights into this past evaluation process are highly valuable in terms of understanding the quality of the evaluation, the extent to which the evaluation met project needs, how the evaluation process and product were used by the project, and the consequences that may have followed any evaluation use. While we acknowledge the potential limitations in your ability to recall every detail about this past evaluation, we ask that you do your best.

If you have any questions as you are taking the survey, feel free to reach out to our team by emailing your questions to Tian Ford at tian@pointedarrows.com.

Thank you,

Pointed Arrows Consulting

1. How satisfied were you with this evaluation process?
    a. Very dissatisfied
    b. Dissatisfied
    c. Somewhat dissatisfied
    d. Somewhat satisfied
    e. Satisfied
    f. Very satisfied
2. Please explain your satisfaction level.
    a. (open-ended)
3. Which MSF-specific principles seemed evident in the conduct of this evaluation? (Example documents that contain specific principles include the *MSF behavioral commitments, SEU ethical guidelines, the Charter, the Chantilly Principles, or La Mancha Agreement)*
    a. (open-ended)
4. How was the evaluation disseminated and used?
    a. (open-ended)
5. Which of the following activities identified in the Use and Dissemination plan were actually enacted? (check all that apply)
    a. Recommendations follow-up
    b. ARO discussions and preparation, round tables, project design
    c. Share the report with DRC and other missions/cells
    d. Evaluation Poster + Report (prepared by SEU) – mail distribution
    e. Short update @ Flash Info/Info Matin
    f. Upload evaluation report on evaluations.msf.org (public)
    g. Presentation to the mission
    h. Presentation and Discussion Session Webinar within OCB
    i. Sharing the report with MOH and partners
    j. Presentation/Discussion with external stakeholders in DRC

6. How satisfied were you with this evaluation’s use and dissemination?
    a. Very dissatisfied
    b. Dissatisfied
    c. Somewhat dissatisfied
    d. Somewhat satisfied
    e. Satisfied
    f. Very satisfied
7. Please explain your satisfaction level.
    a. (open-ended)
8. If this evaluation was used, what were the 1 to 3 most valuable outcomes of that use? Write NA if no use.
    a. (open-ended)
9. What were the most important factors that determined the outcome of this evaluation?
    a. (open-ended)
10. How are evaluation commissioners responsible for the success or failure of evaluations at OCB?
    a. (open-ended)
11. Please share any additional thoughts about the quality or value of this evaluation's processes or products not captured in the previous questions.
    a. (open-ended)

Example Evaluation Manager Block

Our records indicate that you managed this evaluation. The first portion of the survey asks if certain aspects of the Program Evaluation Standards were met. References to the client should be interpreted as the Operational Centre Brussels, including the evaluation commissioner and the project team. The last portion of the survey asks about your experience with the evaluation and for your feedback.

Any questions about the survey can be sent to Tian Ford at tian@pointedarrows.com.

1. Did the evaluation engage evaluators who possessed the needed knowledge, skills, experience, and professional credentials?
    a. Met
    b. Partially met
    c. Partially not met
    d. Not met
    e. I don’t know
    f. Not applicable
2. Did the evaluation engage evaluators whose evaluation qualifications, communication skills, and methodological approach were a good fit to the stakeholders’ situation and needs?
3. Did the evaluation engage evaluators who were appropriately sensitive and responsive to issues of gender, socioeconomic status, race, language, and culture?
4. Did the evaluation engage evaluators who built good working relationships, and listened, observed, clarified, and attended appropriately to stakeholders’ criticisms and suggestions?

5. Did the evaluation engage evaluators who have a record of keeping evaluations moving forward while effectively addressing evaluation users' information needs?
6. Did the evaluation help stakeholders understand the evaluation's boundaries and purposes and engage them to uncover assumptions, interests, values, behaviors, and concerns regarding the program?
7. Did the evaluation determine how stakeholders intend to use the evaluation's findings?
8. Did the evaluation engage the client and stakeholders to weigh stated evaluation purposes—e.g., against their perceptions of dilemmas, quandaries, and desired evaluation outcomes—and to embrace evaluation's bottom line goal of assessing value, e.g., a program's worth, merit, or significance?
9. Did the evaluation help the client group consider possible alternative evaluation purposes, e.g., program planning, development, management, and improvement; program documentation and accountability; and judging the program's quality, impacts, and worth?
10. Did the evaluation engage the client to clarify and prioritize the evaluation's purposes using appropriate tools such as needs assessments and logic models?
11. Did the evaluation engage the client and program stakeholders in an effective process of values clarification, which may include examining the needs of targeted program beneficiaries, the basis for program goals, and the rationale for defined evaluation purposes?
12. Did the evaluation assist the client group to air and discuss their common and discrepant views of what values and purposes should guide the program evaluation?
13. Did the evaluation acknowledge and show respect for stakeholders' possibly diverse perspectives on value matters, e.g., by assisting them to seek consensus or at least reach an accommodation regarding possible alternative interpretations of findings against different values?
14. Did the evaluation clarify the values that would undergird the evaluation, taking account of client, stakeholder, and evaluator positions on this matter?
15. Did the evaluation act to ensure that the client and full range of stakeholders understood and respected the values that would guide the collection, analysis, and interpretation of the evaluation's information?
16. Did the evaluation interview stakeholders to determine their different perspectives, information needs, and views of what constitutes credible, acceptable information?
17. Did the evaluation allow flexibility during the evaluation process for revising the information collection plan pursuant to emergence of new, legitimate information needs?
18. Did the evaluation budget time and resources to allow for meaningful exchange with stakeholders throughout the evaluation process?
19. Did the evaluation engage the full range of stakeholders to assess the original evaluation plan's meaningfulness for the stakeholders intended uses?
20. Did the evaluation regularly obtain stakeholders' reactions to the meaningfulness of evaluation procedures and processes?
21. As appropriate, did the evaluation adapt procedures, processes, and reports to assure that they meaningfully addressed stakeholder needs?
22. Did the evaluation plan to deliver evaluation feedback pursuant to the client group's projection of when they needed reports, while allowing flexibility for responding to changes in the program's timeline and needs?

23. Did the evaluation determine how much technical detail to report by identifying and taking account of the audience's technical background and expectations?
24. In discussing evaluation findings with the client group, did the evaluation stress the importance of applying the findings in accordance with the evaluation's negotiated purposes?
25. Was the evaluation vigilant in identifying, preventing, or appropriately addressing any misuses of the evaluation findings?
26. Did the evaluation include a follow up of evaluation reports to determine if and how stakeholders applied the findings?
27. Did the evaluation prepare a formal management plan including, e.g., the evaluation's goals, procedures, assignments, communication, reporting, schedule, budget, monitoring arrangements, risk management arrangements, and accounting procedures?
28. Did the evaluation recruit evaluation staff members who collectively had the knowledge, skills, and experience required to execute, explain, monitor, and maintain rigor, viability, and credibility in the evaluation process?
29. Did the evaluation systematically oversee and document the evaluation's activities and expenditures?
30. Did the evaluation periodically review the evaluation's progress and, as appropriate, update the evaluation plan and procedures?
31. Did the evaluation assess and confirm the program's evaluability before deciding to proceed with the evaluation?
32. Did the evaluation assure that the selected procedures take account of and equitably accommodate the characteristics and needs of diverse stakeholders?
33. Did the evaluation take into account the interests and needs of stakeholders in the process of designing, contracting for, and staffing the evaluation?
34. Did the evaluation practice even-handedness and responsiveness in relating to all stakeholders, e.g., in the composition of focus groups?
35. Did the evaluation avert or identify and counteract attempts to bias or misapply the findings?
36. Did the evaluation balance effectiveness and efficiency in resource use to help ensure that the evaluation was worth its costs and that sponsors got their money's worth?
37. Did the evaluation document the evaluation's costs, including time, human resources, expenditures, infrastructure support, and foregone opportunities?
38. Did the evaluation plan for and obtain an appropriate approval for needed budgetary modifications over time or because of unexpected problems?
39. Did the evaluation make clear and justify any differential valuing of any stakeholders' evaluation needs over those of others?
40. Did the evaluation carefully monitor and communicate to all right-to-know audiences the evaluation's progress and findings and do so throughout all phases of the evaluation?
41. Did the evaluation scrupulously avoid and prevent any evaluation-related action that was unfair to anyone?
42. Before releasing the evaluation's findings, did the evaluation inform each intended recipient of the evaluation's policies—regarding such matters as right-to-know audiences, human rights, confidentiality, and privacy— and, as appropriate, acquire her or his written agreement to comply with these policies?

43. Did the evaluation provide all right-to-know audiences with access to information on the evaluation's sources of monetary and in-kind support?
44. Was the evaluation frugal in expending evaluation resources?
45. Did the evaluation employ professionally accepted accounting and auditing practices?
46. Did the evaluation maintain accurate and clear fiscal records detailing exact expenditures, including adequate personnel records concerning job allocations and time spent on the job?
47. Did the evaluation make accounting records and audit reports available for oversight purposes and inspection by stakeholders?
48. Did the evaluation assure that the evaluation team included or had access to expertise needed to investigate the applicable types of reliability?
49. Did the evaluation ensure that the collection of information was systematic, replicable, adequately free of mistakes, and well documented?
50. Did the evaluation establish and implement protocols for quality control of the collection, validation, storage, and retrieval of evaluation information?
51. Did the evaluation document and maintain both the original and processed versions of obtained information?
52. Did the evaluation retain the original and analyzed forms of information as long as authorized users needed it?
53. Did the evaluation store the evaluative information in ways that prevented direct and indirect alterations, distortions, destruction, or decay?
54. Did the evaluation plan for specific procedures to avert and check for threats to reaching defensible conclusions, including analysis of factors of contextual complexity, examination of the sufficiency and validity of obtained information, checking on the plausibility of assumptions underlying the evaluation design, and assessment of the plausibility of alternative interpretations and conclusions?
55. Did the evaluation consistently check and correct draft reports to assure they were impartial, objective, free from bias, responsive to contracted evaluation questions, accurate, free of ambiguity, understood by key stakeholders, and edited for clarity?
56. How satisfied were you with this evaluation process?
    a. Very dissatisfied
    b. Dissatisfied
    c. Somewhat dissatisfied
    d. Somewhat satisfied
    e. Satisfied
    f. Very satisfied
57. Please explain your satisfaction level.
    a. (open-ended)

58. Which MSF-specific principles seemed evident in the conduct of this evaluation? (Example documents that contain specific principles include the *MSF behavioral commitments, SEU ethical guidelines, the Charter, the Chantilly Principles, or La Mancha Agreement)*
    a. (open-ended)
59. How was the evaluation disseminated and used?
    a. (open-ended)

60. Which of the following activities identified in the Use and Dissemination plan were actually enacted? (check all that apply)
    a. Recommendations follow-up
    b. ARO discussions and preparation, round tables, project design
    c. Share the report with DRC and other missions/cells
    d. Evaluation Poster + Report (prepared by SEU) – mail distribution
    e. Short update @ Flash Info/Info Matin
    f. Upload evaluation report on evaluations.msf.org (public)
    g. Presentation to the mission
    h. Presentation and Discussion Session Webinar within OCB
    i. Sharing the report with MOH and partners
    j. Presentation/Discussion with external stakeholders in DRC
61. How satisfied were you with this evaluation's use and dissemination?
    a. Very dissatisfied
    b. Dissatisfied
    c. Somewhat dissatisfied
    d. Somewhat satisfied
    e. Satisfied
    f. Very satisfied
62. Please explain your satisfaction level.
    a. (open-ended)
63. If this evaluation was used, what were the 1 to 3 most valuable outcomes of that use? Write NA if no use.
    a. (open-ended)
64. What were the most important factors that determined the outcome of this evaluation?
    a. (open-ended)
65. How are evaluation managers responsible for the success or failure of evaluations at OCB?
    a. (open-ended)
66. Please share any additional thoughts about the quality or value of this evaluation's processes or products not captured in the previous questions.
    a. (open-ended)

EXAMPLE EVALUATOR BLOCK

Our records indicate that you conducted this evaluation. The first portion of the survey asks if certain aspects of the Program Evaluation Standards were met. References to the client should be interpreted as the Operational Centre Brussels, including the evaluation commissioner and the project team. The last portion of the survey asks about your experience with the evaluation and for your feedback.

Any questions about the survey can be sent to Tian Ford at tian@pointedarrows.com.

1. Did the evaluation engage and serve the full range of stakeholders in an even-handed manner, regardless of their politics, personal characteristics, status, or power?

   a. Met
   b. Partially met
   c. Partially not met
   d. Not met
   e. I don't know
   f. Not applicable

2. Was the evaluation open to and did it thoughtfully consider stakeholders' contradictory views, interests, and beliefs regarding the program's prior history, goals, status, achievements, and significance?
3. Did the evaluation avert or counteract moves by powerful stakeholders to dominate in determining evaluation purposes, questions, and procedures and interpreting outcomes?
4. Did the evaluation revisit agreements over time and negotiate revisions as appropriate?
5. Did the evaluation adhere to applicable federal, state, local, and tribal regulations and requirements, including those of Institutional Review Boards, local/tribal constituencies, and ethics committees that authorize consent for conduct of research and evaluation studies?
6. Did the evaluation take the initiative to learn, understand, and respect stakeholders' cultural and social backgrounds, local mores, and institutional protocols?
7. Did the evaluation monitor the interactions of evaluation team members and stakeholders and act as appropriate to ensure continuing, functional, and respectful communication and interpersonal contacts throughout the evaluation?
8. How satisfied were you with this evaluation process?
   a. Very dissatisfied
   b. Dissatisfied
   c. Somewhat dissatisfied
   d. Somewhat satisfied
   e. Satisfied
   f. Very satisfied
9. Please explain your satisfaction level.
   a. (open-ended)
10. How was the evaluation disseminated and used?
   a. (open-ended)
11. What were the most important factors that determined the outcome of this evaluation?
   a. (open-ended)
12. Please share any additional thoughts about the quality or value of this evaluation's processes or products not captured in the previous questions.
   a. (open-ended)
13. Given your experience with this evaluation, please provide feedback for the Stockholm Evaluation Unit to consider for improving future evaluation processes.
   a. (open-ended)

ANNEX XI: LIST OF KEY INFORMANTS AND FOCUS GROUP PARTICIPANTS BY ROLE

OCB Department	Relation to SEU	ME Role	Method
OCB HQ	Consultation Group/Steering Committee Member	Co-commissioner	Email Interview
OCB HQ	Consultation Group/Steering Committee Member	Co-commissioner	Interview
OCB HQ	ME Consultation Group Member	Consultation Group Member	Interview
Field-based (operations)	ME Consultation Group Member	Consultation Group Member	Interview
Field-based (operations)	ME Consultation Group Member	Consultation Group Member	Email Interview
Field-based (operations)	ME Consultation Group Member	Consultation Group Member	Email Interview
OCB HQ	ME Consultation Group Member	Consultation Group Member	Interview
OCB HQ	Steering Committee Member	Primary User	Interview
OCB HQ	Steering Committee Member	Primary User	Interview
OCB HQ	Steering Committee Member	Primary User	Email Interview
MSF Sweden	Steering Committee Member/Acting Chairperson	Primary User	Interview
MSF Sweden	Steering Committee Member	Primary User	Email Interview
Evaluation Unit	SEU Team Member	Primary User	Interview
Evaluation Unit	SEU Team Member	Primary User	Interview
Evaluation Unit	SEU Team Member	Primary User	Interview
Evaluation Unit	SEU Team Member	Primary User	Interview
Evaluation Unit	SEU Team Member	Focal Point/Manager/ Primary User	Interview
Medical Department	External	Informant/secondary user	Focus Group
Medical Department	External	Informant/secondary user	Focus Group
Medical Department	External	Informant/secondary user	Focus Group
Medical Department	External	Informant/secondary user	Focus Group
Analytical Department	External	Informant/secondary user	Focus Group
Analytical Department	External	Informant/secondary user	Focus Group
Analytical Department	External	Informant/secondary user	Focus Group
Medical Department	External	Informant/secondary user	Email Interview
GD Direction	External	Informant/secondary user	Interview

ANNEX XII: TABLE OF SURVEY RESPONDENTS BY EVALUATION CASE AND ROLE

Evaluation	Evaluator	Evaluation Manager	Project Contact	Evaluation Commissioner	Total
ARCHE	1	1	0	0	2
BILIC	1	1	0	0	2
BOLIM	1	1	0	1	3
BUDGE	0	0	0	0	0
COMME	1	1	0	0	2
DGDFM	1	1	0	1	3
DIGHP	0	1	1	1	3
EBOLA	2	1	0	0	3
EMRKS	0	0	0	0	0
EPOOL	2	1	0	0	3
ESHIV	1	1	0	0	2
FRCOH	0	1	1	0	2
GUCCE	0	1	1	1	3
HIVKIN	1	1	0	0	2
HREVA	0	1	0	0	1
IDAII	1	1	0	1	3
MASTE	0	1	0	0	1
MAURT	0	0	0	0	0
MBADO	1	1	0	1	3
MUMPO	1	1	0	0	2
MVGCE	0	0	0	1	1
NCDKE	0	1	1	0	2
OCBFE	0	0	1	0	1
OCBPR	0	0	0	1	1
OCHMU	0	1	0	0	1
REACH	1	1	1	0	3
SUPCH	1	1	1	0	3
EVAL21	1	1	0	0	2
USCOV	1	0	0	0	1
VOTTR	0	1	0	0	1
VTCAR	0	1	0	0	1
Total Responses	18	24	7	8	57
Total Possible	42	34	30	31	137
Response Rate	43%	76%	23%	23%	41%

ANNEX XIII: PRGES AND ALNAP PROFORMA CHECKLIST DASHBOARD CODEBOOK

PORTFOLIO-WIDE DECISION RULES, INTERPRETATION PRINCIPLES, AND ASSUMPTIONS

Minimum standards: checkpoints will be deemed met if the actual evaluation product or process is mostly embodied in the standard definition, even if there could be room for improvement of the actual product or process.

Letter of the standard versus spirit of the standard: if it appears the aim of a standard is met, but in a slightly different way than described in the standard, this will be determined sufficient. For example, if a standard asks for a final report template in a ToR, and it is shared elsewhere, this will be sufficient to meet the spirit of the standard.

Evidence for standards can come from multiple and more than one source: some standards may suggest the document source of a required piece of information for a standard to be met. If the product standard is evident in any of the evaluation documents, it should be deemed met.

Treatment of multiple elements: some standards list multiple elements to be present in an evaluation process or product, such as the avoidance of discrimination of evaluation participants based on multiple identity markers. Where any one of those listed factors is not met, the whole checkpoint is deemed not met. Other standards list multiple elements that comprise a standard, but may not all need to be present to be met. This is especially the case where the list of elements seem more illustrative, (for instance where standards offer a list after "e.g") reviewers can, using the *spirit of the standard* principle, determine a standard is met, even if one element may not be manifest or manifest as it is specifically described.

Support all judgments with evidence: each decision on whether a standard is or isn't met, unqualified by an asterisk, will be supported by the document or policy source of the evidence for that standard.

Process versus product standards: some standards refer to the absence or presence of elements in evaluation documents and others for qualities in processes. Product standards are either met or not met in documents or not met in their absence in documentation. Process standards can be evident in documentation, but not always. When not initially evident in documentation, process-related standards that are unclear if they are met will be asked of evaluators and evaluation managers in an online survey.

Absence of evidence for a process standard will be treated as the lack of that standard: when survey respondents don't know if a standard was met or not, they will be marked as not met. If no response is received from evaluators or managers for certain standards, the lack of evidence that a standard was met or not will be interpreted as not being met, and indicated with an asterisk.

Not applicable standards: despite the general nature of these standards, it is expected that some may not be applicable by the nature and purpose distinction of the evaluations being assessed. Standards that are deemed not applicable will be indicated as "met*" and will be treated the same as "met" standards. The number of not applicable standards will be reported for each evaluation and for the portfolio overall.

Unusually rare instances of evaluation malpractice will be assumed not applicable: some exceptional standards were deemed not applicable by the extraordinariness of their nature such as evaluation sabotage in U8CP2 or the intentional biasing or misapplication of findings in F3CP5. The responsibility to indicate if these issues were operational rests with the members of the SEU.

Evaluation standards can be systematically met based on prior knowledge: if a process or product standard is covered by an existing SEU policy, the meta-evaluation team will assume this standard is met. For example, multiple checkpoints will be met with the assumption that all evaluations had the formation of a consultation group per SEU policy or had bi-weekly meetings between evaluators and managers, per evaluation contracts. Instances of these inferences have been indicated in the comments about each standard. The SEU will have the responsibility to correct any exceptions to these standard operating procedures or policies where these assumptions were not met for any specific evaluation case.

Appropriate interpretations: many standards ask for certain products to be shared with “all right to know audiences” or processes to include the “full range of stakeholders.” The assumption of consultation group representativeness, public evaluation dissemination, and or a list of additional key informants consulted in the inception phase will meet this specific clause in standards that include this language and documentation.

STANDARD-SPECIFIC DECISION RULES, INTERPRETATION PRINCIPLES, AND ASSUMPTIONS

ALNAP PROFORMA CHECKLIST

Applicability of OECD-DAC criteria in ALNAP Proforma 4.3: if an evaluation does not include one or more OECD-DAC criteria, those omitted criteria will be treated as not relevant for the specific evaluation, even if they could theoretically be investigated. Evaluations will only receive “not met” ratings for these criteria standards if they scoped criteria and had insufficient evidence to make a conclusion, or where conclusions were judged insufficiently warranted by evidence presented.

Applicability of cross-cutting themes in ALNAP Proforma 4.4: as opposed to section 4.3, these standards will be assumed applicable, despite scoping, unless logically not applicable on a case-by-case basis. That is to say, if an evaluation didn’t include a gender analysis (or other cross-cutting themes) by design, it will not be considered “not applicable” by default, but assume applicability as the rule, and look for exceptions.

ALNAP 1.2, “The TOR should clarify the commissioning agency‘s expectation of good humanitarian evaluation practice. (e.g., application of DAC criteria;4 reference to international standards including international law; multi-method approach i.e., quantitative and qualitative; consultation with key stakeholders to inform findings, conclusions and recommendations; and gender analysis).” After initial mixed interpretation by raters, it was determined that this standard is met systematically across the portfolio in spirit by the following evidence: almost every ToR references OECD-DAC; SEU guiding documents including the manifesto, ethical guidelines, and framework articulate what this means; there is an onboarding meeting where these values are likely re-iterated, if these documents were not shared nor consultants expected to sign; we and likely many other consultants were required to familiarize themselves and sign their name in following the ethical guidelines and possibly other documents.

PROGRAM EVALUATION META-EVALUATION CHECKLIST

Evaluation Accountability Section 2 Checkpoints 1-6 (E2CP1-6): after we initially decided to systematically "not meet" these standards, following the "spirit of the standard" principle, we have determined the SEU has processes in place to determine quality through systematic feedback loops, decision gates, medical and technical referents when applicable. Procedures are described in the section Quality Assurance and Control. This is effectively meta-evaluation under a different name. After further consideration, E2CP1-3 are being systematically met under the "spirit" of these standards, though they and certainly the other checkpoints in this sub-criterion could likely be improved through more formalized meta-evaluation processes.

E3CP1-6: all of these checkpoints that pertain to external meta-evaluation have been deemed systematically met across all evaluation cases with the commission of this external portfolio meta-evaluation, even though no known external meta-evaluations were conducted for individual evaluations.

F4CP6: "Document the evaluation's benefits, including contributions to program improvement, future funding, better informed stakeholders, and dissemination of effective services" after an initial interpretation that this standard could be prospective in inception reports or final reports, we determined the original intent was a retrospective documentation and accordingly decided to systematically "umeet" this checkpoint unless there were select instances in management responses of evaluation consequences documented. We found no such evidence across the portfolio.

U7CP5: "Plan and budget evaluation follow-up activities so that the evaluator can assist the client group to interpret and make effective use of the final evaluation report" after initially systematically not meeting this based on absence of line items in budget and not interpreting final oral presentations as sufficient, under the spirit of the standard principle, finally interpret this checkpoint sufficiently met across the board with the coordinated efforts in use and dissemination plans and other wrap-around and follow-up support provided by evaluation managers, coordinator, and head of unit, as opposed to interpreting solely coming from the external evaluators.

A1CP4 "Identify the persons who determined the evaluation's conclusions, e.g., the evaluator using the obtained information plus inputs from a broad range of stakeholders" After initially meeting most if not all of these checkpoints across evaluation cases based on an assumption that the conclusions come solely from the evaluators. We determined this seems to be a high burden of proof and is explicit checkpoint. There are some instances where findings from informants and survey respondents are shared in terms of differing views about questions, or at least one report that talked about the difference between initial recommendations and recommendations revised or co-created with SEU. But this instance about recommendations occurred only once and certainly not with conclusions. Therefore, we are systematically not meeting this checkpoint.

F2CP1: "Assess and confirm the program's evaluability before deciding to proceed with the evaluation" from analysis of SEU procedures, artifact review, evidence of low evaluability in not being able to answer some key evaluation questions, and discussion with head of unit, determined to systematically not meet this checkpoint, except for instances with evaluations that have scoping documents for 8 evaluations.

U5CP1 "Interview stakeholders to determine their different perspectives, information needs, and views of what constitutes credible, acceptable information" After an initial strict interpretation of this checkpoint in needing to reference all elements, especially discussions of credibility of evidence, which resulted in low meeting rates, we reconsidered our position and determined this standard is met by the "spirit of the standard" in that most if not all evaluations have long inception stages where they meet with at least consultation group members, amid others, and while credibility of evidence may not be addressed, evidence suggest getting sense of information needs, perspectives and acceptable information are likely discussed. We decided to systematically meet this across the portfolio.

U6CP3: "During the evaluation process, regularly visit with stakeholders' to assess their evaluation needs and expectations, also, as appropriate, to obtain their assistance in executing the evaluation plan" we are systematically meeting this with assumptions of clauses in contract that require bi-weekly meeting, even though there may be variation in this across managers, as suggested by head of unit.

A8CP3: "Schedule formal and informal reporting in consideration of user needs, including follow-up assistance for applying findings" we are "meeting" systematically based on contractual obligations, CG formation assumption, and specifically the second clause being met with SEU coordinator/manager follow-up wrap-around support.

A8CP4: "Employ multiple reporting mechanisms, e.g., slides, dramatizations, photographs, PowerPoint©, focus groups, printed reports, oral presentations, telephone conversations, and memos" we are systematically meeting where there are more than one reporting documents. Strong evidence against these policies and products happening or existing to determine "not met"

F3CP1 "Investigate the program's cultural, political, and economic contexts by reviewing such items as the program's funding proposal, budget documents, organizational charts, reports, and news media accounts and by interviewing such stakeholders as the program's funder, policy board members, director, staff, recipients, and area residents" systematically meeting based on lengthy inception stage, in-depth inception reports, consultation group interviews.

F4CP1: "Negotiate a budget--ensuring that the contracted evaluation work can be completed efficiently and effectively—to include the needed funds and the necessary in-kind support and cooperation of program personnel" in instances where no artifacts existed, this checkpoint was systematically met under the assumption these conversations had to have taken place for contracting.

P7CP1: "Plan and obtain approval of the evaluation budget before beginning evaluation implementation" in the few instances where budget artifacts were not available, we actually met these following the same decision rule for F4CP1.

A8CP1: "Reach a formal agreement that the evaluator will retain editorial authority over reports" Initially, all evals were docked by not having evaluation final editorial authority, assuming the clause in the contract about co-authorship negated that, but we determined this may have pertained more to rights of report ownership. Many evaluation title pages, but not all, have disclaimers that these were prepared independently and views may not be shared by MSF. Which suggests editorial authority, as well as the SEU ethical guidelines have independence, neutrality, impartiality discussed. We decided to systematically meet this under the spirit of the standard given a host of indications that SEU views evaluators have editorial authority, especially given management responses are a place to agree or disagree, and that this is transparent, and not something requested and assented to by default in report writing.

U3CP5 "Provide for engaging the client group periodically to revisit and, as appropriate, update the evaluation's purposes" though it is difficult to determine last clause in this checkpoint, we are systematically meeting this standard with the knowledge of a clause in the contract that requires bi-weekly meetings, unless indicated otherwise in the surveys that had this item included.

A3CP6 "Examine and discuss the consistency of scoring, categorization, and coding and between different sets of information, e.g., assessments by different observers" Initially rated as not met or not applicable due to only one evaluator, but later recognized this could pertain to intra-rater reliability across sites, documents, groups and decided to only meet if this was explicitly discussed, regardless of the number of evaluators.

P3CP3 "Make clear to the client and stakeholders the evaluator's ethical principles and codes of professional conduct, including the standards of the Joint Committee on Standards for Educational Evaluation" Initially required evaluations to reference JCSEE, but later recognized this could be met with any reference to ethical or quality frameworks or criteria and that this was an instance of a standard including a reference for an example, or at least should be as the JCSEE are not the only quality/ethical framework in the business.

P2CP6 – "Revisit evaluation agreements over time and negotiate revisions as appropriate" After initial differences of interpretation by raters caught by reliability checks, it was determined to systematically meet this one, based on the facts that contract amendments were common practice as needed. Missing evidence for surveys that were sent out affected reliability too much, in that these were determined not met, and no checkpoint was indicated as not met with evidence of not meeting, so systematically met with combination of evidence of amendments, and clauses in contracts, and no instance of evidence of this not happening, merely lack of evidence adversely affecting reliability.

U8CP5: this item had survey data construction error in failing to include it as an item in all surveys that needed it. It has been coded at "met*" across the board in cases of missing data and with "met" in all cases of survey responses, as all survey responses for this item that we had were "met"

U6CP6: this item had survey data construction error in failing to include it as an item in all surveys that needed it. It has been coded at "met*" across the board in cases of missing data unless otherwise indicated in the few evaluations that had this item on their survey. Only one evaluation survey respondent indicated this was not met.

>-<

ANNEX XIV: RECOMMENDATION SUPPLEMENTAL RESOURCES

PrgES

Recommendation Code	Resource
Feasibility	Evaluability resources: ● Planning Evaluability Assessments: A Synthesis of the Literature with Recommendations, Rick Davies ● Evaluability Checklist, Better Evaluations ● UNICEF Evaluability Assessment Guidance
Evaluation Accountability	Principles-Focused Evaluation: ● Principles-Focused Evaluation: The GUIDE ● BetterEvaluation P-FE Entry Culturally Responsive and Equitable Evaluation (CREE): ● CULTURALLY RESPONSIVE EVALUATION Theory, Practice, and Future Implications ● Center for Culturally Response Evaluation and Assessment Publication List ● Equitable Evaluation Initiative ● Cultural Reading of the 2nd Edition of the Program Evaluation Standards, AEA ● Call to Action Series, FEAN ● Righting Systemic Wrongs Organizational Self-Assessment ● Evaluation is SO White: Systemic Wrongs Reinforced by Common Practices and How to Start Righting Them, Fontane Lo & Rachele Espiritu

UNEG Norms and Standards

Recommendation Code	Resource
Professionalism	● The AEA Competencies Framework
Human Rights and Gender Equality	● Integrating Human Rights and Gender Equality in Evaluations, UNEG ● UNEG Repository of Guidance, Analysis, and Good Practice Resources on Intergrating Gender Equality and Human Rights in Evaluations ● See PEMC, “Evaluation Accountability” resources on Culturally Responsive and Equitable Evaluation (CREE). ● Cultural Reading of the 2nd Edition of the Program Evaluation Standards, AEA

	• Call to Action Series, FEAN • Righting Systemic Wrongs Organizational Self-Assessment
Ethics	• UNEG Ethical Guidelines for Evaluation
Evaluation Guidelines	• See PEMC, "Evaluation Accountability" resources on principles focused evaluation.
Evaluability Assessments	• See PEMC, "Feasibility" resources on evaluability assessments.
Terms of Reference	• Evaluation is SO White: Systemic Wrongs Reinforced by Common Practices and How to Start Righting Them, Fontane Lo & Rachele Espiritu
Evaluation Reports and Products	General Logic of Evaluation: • Fournier, D. M. (1995). Establishing evaluative conclusions: A distinction between general and working logic. New Directions for Evaluation, 1995(68), 15–32. doi:10.1002/ev.1017 • Scriven, M. (1995). The logic of evaluation and evaluation practice. New Directions for Evaluation, 1995(68), 49–70. doi:10.1002/ev.1019 • Scriven, M. (2007) The Logic of Evaluation • Stake, R., Migotsky, C., Davis, R., Cisneros, E. J., Depaul, G., Dunbar, C., … Chaves, I. (1997). The Evolving Syntheses of Program Value. American Journal of Evaluation, 18(1), 89–103. doi:10.1177/109821409701800110 • Ozeki, S., Coryn, C. L. S., & Schröter, D. C. (2019). Evaluation logic in practice. Evaluation and Program Planning, 76, 101681. doi:10.1016/j.evalprogplan.2019.101681 • Gullickson, A. M. (2020). The whole elephant: Defining evaluation. Evaluation and Program Planning, 79, 101787. doi:10.1016/j.evalprogplan.2020.101787 • Montrosse-Moorhead, B. (2021, November 24). Evaluating items commonly found in a home using evaluation logic (Version 2.0).
Follow-up	• UNICEF Evaluation Management Response

EMQF

Recommendation Code	Resource
Evaluation Report and products	• See UNEG, "Evaluation Reports and Products" resources on general evaluation logic.
Evaluability Assessment	• See PEMC, "Feasibility" resources on evaluability assessments.

Choosing Criteria	• Applying Evaluation Criteria Thoughtfully, OECD-DAC • Evaluation Criteria for Evaluating Transformation: Implications for the Coronavirus Pandemic and the Global Climate Emergency, Micheal Q. Patton • Teasdale, R. M. (2021). Evaluative Criteria: An Integrated Model of Domains and Sources. American Journal of Evaluation, 42(3), 354–376. doi:10.1177/1098214020955226
Annual Report	• See PEMC, "Evaluation Accountability" resources on principles focused evaluation.
Re-coding of PrgES and ALNAP Indicators	• The Evaluation Theory Tree: • Mertens & Wilson (2019) Chapter Two of Program Evaluation Theory and Practice 3rd Ed. Mertens, D.M. (2015) PHILOSOPHICAL ASSUMPTIONS AND PROGRAM EVALUATION

Miscellaneous Resources:

Report Section	Resource
Evaluation Use and Use Outcomes	• Alkin, M. C., & King, J. A. (2016). The Historical Development of Evaluation Use. American Journal of Evaluation, 37(4), 568–579. doi:10.1177/1098214016665164 • Alkin, M. C., & King, J. A. (2017). Definitions of Evaluation Use and Misuse, Evaluation Influence, and Factors Affecting Use. American Journal of Evaluation, 38(3), 434–450. doi:10.1177/1098214017717015 • King, J. A., & Alkin, M. C. (2018). The Centrality of Use: Theories of Evaluation Use and Influence and Thoughts on the First 50 Years of Use Research. American Journal of Evaluation, 109821401879632. doi:10.1177/1098214018796328 • Patton, M. Q. (2020). Evaluation Use Theory, Practice, and Future Research: Reflections on the Alkin and King AJE Series. American Journal of Evaluation, 109821402091949. doi:10.1177/1098214020919498
Recommendations	• Evaluation Center Eval Café Presentation from Lori Wingate about Recommendations in Evaluation

>-<

ANNEX XV: EVALUATION CASE ABBREVIATIONS, TITLES, AND SUMMARY

Evaluation Code	Title	Year	Coverage
ARCHE	The Arche Project: Centre of Traumatology	2021	Burundi
BILIC	MSF-OCB's Malaria Project	2020	DRC
BOLIM	Maternal and Child Sexual and Reproductive Health Intervention	2021	Bolivia
BUDGE	OCB Operational "Overspend"	2018	Belgium
COMME	MSF-OCB's Corridor Programs for Key Populations	2018	SnA Africa
DGDFM	Optimizing HIV, TB, & NCD, Treatment in 5 Sub-Saharan Africa Countries	2017	Southern Africa
DIGHP	COVID-19's Digital Health Promotion	2021	Various
EBOLA	MSF-OCB's Ebola Intervention	2020	DRC
EMRKS	The EMR Deployment in Kabinda VIH Hospital	2018	DRC
EPOOL	MSF-OCB's Hurricane Matthews Emergency Response	2017	Haiti
ESHIV	The Eshowe HIV Project	2021	South Africa
FRCOH	MSF-OCB's Field Recentralization Monitoring Exercise	2021	South Africa
GUCCE	MSF's Cervical Cancer Intervention	2021	Zimbabwe
HIVKIN	The HIV Decentralization Initiative	2020	DRC
HREVA	Data Drives Design: Moving Forward with Implementing the OCB Evaluation Process	2017	Various
IDAII	MSF Emergency Response to Cyclone IDAI	2019	Mozambique
MASTE	MSF Rural Health Services Scholarship Programme	2018	Malawi
MAURT	Assistance to Malian refugees and resident population	2018	Mauritania
MBADO	Adolescents Sexual and Reproductive Health Project	2021	Zimbabwe
MUMPO	The Catalytic Role of Mumbai Project with Regards to Policy Changes	2021	India
MVGCE	Real Time Evaluation of a Measles Vaccination Campaign in Conakry City	2017	Guinea
NCDKE	Clinical Mentoring in MSF's Non-Communicable Disease Project	2020	Kenya

OCBFE	MSF OCB's Field Opportunity Envelope Review	2017	Belgium
OCBPR	OCB Operational Prospects 2014-2017 Review	2017	Belgium
OCHMU	An Organizational Assessment of OCB-MSF's Hospital Management Unity	2020	Belgium
REACH	MSF Reaction Assessment Collaboration Hub: The Reach Project	2020	Hong Kong
SUPCH	MSF-OCB's End-to-End Supply Chain	2021	Global
EVAL21	MSF-OCB's Project	2021	Middle East
USCOV	MSF-USA's COVID-19 Evaluation of Seven Projects	2021	US
VOTTR	MSF's Treatment & Rehabilitation of Victims of Torture Programs	2020	Egypt, Italy, Greece
VTCAR	MSF-OCB Torture Rehabilitation Project	2017	Mediterranean

>-<

ANNEX XVI: META-META-EVALUATION ATTESTATION

This meta-evaluation addressed four major merit and worth questions about the quality and value of 31 SEU evaluation cases and the system that implemented them. It answered them through the application of the PEMC, ALNAP Proforma, UNEG Standards and Norms, and the EMQF. Assessing the quality of evaluation systems is a professional imperative and it would be hubris to sit in the position of judgement and not apply the same sort of judgment to one's own practice. Especially when the reviewing itself provides such a rich opportunity for reflection and the potential for improvement. Publishing it here is our effort at transparency. We used the Program Evaluation Meta-evaluation Checklist (PEMC) to rate this meta-evaluation and found it to be Excellent with a score of 97%. The ratings were assigned through consensus of the meta-evaluation team. The meta-evaluation met almost all indicators (n = 175, 97%), with very few not applicable (n = 19, 9%). Those not met were 2 indicators in the accuracy criterion, 2 in the utility criterion, and 1 feasibility criterion indicator. Indicators that were not met:

- Utility sub-criteria 1, evaluator credibility, checkpoint 6: Give stakeholders information on the evaluation plan's technical quality and practicality, e.g., as assessed by an independent evaluation expert
- Utility sub-criteria 2, attention to stakeholders, checkpoint 3: Search out & invite input from groups or communities whose perspectives are typically excluded, especially stakeholders who might be hindered by the evaluation
- Feasibility sub-criteria 3, contextual viability, checkpoint 1: Investigate the program's cultural, political, and economic contexts by reviewing such items as the program's funding proposal, budget documents, organizational charts, reports, and news media accounts and by interviewing such stakeholders as the program's funder, policy board members, director, staff, recipients, and area residents
- Accuracy sub-criteria 8, communicating and reporting, checkpoint 1: Reach a formal agreement that the evaluator will retain editorial authority over reports
- Accuracy sub-criteria #8, communicating and reporting, checkpoint 2: Reach a formal agreement defining right-to-know audiences and guaranteeing appropriate levels of openness and transparency in releasing and disseminating evaluation findings

The table below provides the scores and ratings for each sub-criteria, criteria, and combined total.

THE UTILITY STANDARDS ARE INTENDED TO ENSURE THAT AN EVALUATION IS ALIGNED WITH STAKEHOLDERS' NEEDS SUCH THAT PROCESS USES, FINDINGS USES, AND OTHER APPROPRIATE INFLUENCES ARE POSSIBLE.

U1 Evaluator Credibility.		U2 Attention to Stakeholders.		U3 Negotiated Purposes.		U4 Explicit Values.		U5 Relevant Information.		U6 Meaningful Processes and Products.		U7 Timeliness and Appropriate Communication and Reporting.		U8 Concern for Consequences and Influence.		Total Score	Total Rating
Score	Rate	Score	Rate	Score	Rate	Score	Rate	Score	Rate	Score	Rate	Score	Rate	Score	Rate		
83%	Very Good	83%	Very Good	100%	Excellent	100%	Excellent	100%	Excellent	100%	Excellent	100%	Excellent	100%	Excellent	96%	Excellent

THE FEASIBILITY STANDARDS ARE INTENDED TO ENSURE THAT AN EVALUATION IS VIABLE, REALISTIC, CONTEXTUALLY SENSITIVE, RESPONSIVE, PRUDENT, DIPLOMATIC, POLITICALLY VIABLE, EFFICIENT, AND COST EFFECTIVE.

F1 Project Management.		F2 Practical Procedures.		F3 Contextual Viability.		F4 Resource Use.		Total Score	Total Rating
Score	Rate	Score	Rate	Score	Rate	Score	Rate		
100%	Excellent	100%	Excellent	83%	Very Good	100%	Excellent	95%	Excellent

THE PROPRIETY STANDARDS ARE INTENDED TO ENSURE THAT AN EVALUATION WILL BE CONDUCTED PROPERLY, FAIRLY, LEGALLY, ETHICALLY, AND JUSTLY WITH RESPECT TO (1) EVALUATORS' AND STAKEHOLDERS' ETHICAL RIGHTS, RESPONSIBBILITIES, AND DUTIES; (2) SYSTEMS OF RELEVANT LAWS, REGULATIONS, AND RULES; AND (3) ROLES AND DUTIES OF PROFESSIONAL EVALUATORS.

P1 Responsive and Inclusive Orientation.		P2 Formal Agreements.		P3 Human Rights and Respect.		P4 Clarity and Fairness.		P5 Transparency and Disclosure.		P6 Conflicts of Interests.		P7 Fiscal Responsibility.		Total Score	Total Rating
Score	Rate	Score	Rate	Score	Rate	Score	Rate	Score	Rate	Score	Rate	Score	Rate		
100%	Excellent	100%	Excellent	100%	Excellent	100%	Excellent	100%	Excellent	100%	Excellent	100%	Excellent	100%	Excellent

THE ACCURACY STANDARDS ARE INTENDED TO ENSURE THAT AN EVALUATION EMPLOYS SOUND THEORY, DESIGNS, METHODS, AND REASONING IN ORDER TO MINIMIZE INCONSISTENCIES, DISTORTIONS, AND MISCONCEPTIONS AND PRODUCE AND REPORT TRUTHFUL EVALUATION FINDINGS AND CONCLUSIONS.

A1 Justified Conclusions and Decisions	A2 Valid Information.	A3 Reliable Information.	A4 Explicit Program and Context Descriptions.	A5 Information Management.	A6 Sound Designs and Analyses.	A7 Explicit Evaluation Reasoning.	A8 Communicating and Reporting.	Total Score	Total Rating

Score	Rate	Score	Rate	Score	Rate	Score	Rate	Score	Rate	Score	Rate	Score	Rate	Score	Rate		
100%	Excellent	100%	Excellent	100%	Excellent	100%	Excellent	100%	Excellent	100%	Excellent	100%	Excellent	67%	Good	95%	Excellent
THE EVALUATION ACScoreABILITY STANDARDS ARE INTENDED TO ENSURE THAT AN EVALUATION IS SYSTEMATICALLY, THOROUGHLY, AND TRANSPARENTLY DOCUMENTED AND THEN ASSESSED, BOTH INTERNALLY AND EXTERNALLY FOR ITS UTILITY, FEASIBILITY, PROPRIETY, AND ACCURACY.																	
E1 Evaluation Documentation.		E2 Internal Metaevaluation.		E3 External Metaevaluation.												Total Score	Total Rating
Score	Rate	Score	Rate	Score	Rate												
100%	Excellent	100%	Excellent	100%	Excellent											100%	Excellent
																Total Score	Total Rating
																97%	Excellent

>-<

ANNEX XVII: METAE MANAGEMENT RESPONSE

As the meta-evaluation team, we recommend drafting and including a management response for this meta-evaluation as an annex in the public version. While management responses have taken various shapes in the past at OCB, we suggest refreshing the process and product after some consideration for what makes sense. The response format to this meta-evaluation does not need to be the final form, nor does reconfiguring an updated response template delay some sort of response by management in writing to be annexed to this public meta-evaluation report. An updated management response process and product would be a key component of a refreshed evaluation follow-up process. Some ideas for an updated response are:

- The use of some of the most important meta-evaluation criteria that were used in this study
- Standardizing the application of those criteria using a rubric with standards
- An opportunity to have multiple intended user group representatives comment about their experience or judgement of the evaluation quality, such as the evaluation manager, the commissioner, and even project staff.
- Responses from management on the extent to which they agree with main conclusions
- Responses from management on the extent to which they agree with main recommendations, and if agreed, what are the planned course of action

There are multiple management response templates and guidance, mostly from the UN family of evaluation units. This is likely the best guidance document that would need some adaptation for the SEU/OCB context.

>-<

Stockholm Evaluation Unit
http://evaluation.msf.org/
Médecins Sans Frontières

Independently written by
Michael Harnar, Zach Tilton and Tian K. Ford.
(December 2022)