Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs,
Education and Research EAER
State Secretariat for Economic Affairs SECO

SYS
PONS

REPORT

PERFORMANCE & QUALITY REPORT 2023 - 2024

Assessment of Quality and Performance of SECO Economic Cooperation and Development's Evaluated Portfolio

SEPTEMBER 2025

AUTHORS

State Secretariat for Economic Affairs (SECO)
Economic Cooperation and Development Division
Evaluation Unit (WEQA)

Johannes Schneider,
Nicola Malacarne

we.evaluation@seco.admin.ch

Syspons GmbH

Lennart Raetzell,
Lena Häberlein,
Dr. Matthias Stelter,
Julia Forke

info@syspons.com

Foreword

The 4th Performance and Quality Report provides an important opportunity to review our achievements and plan the way forward. Every two years, we evaluate the performance of our economic development cooperation. The latest review confirms that we are on the right track, while also highlighting areas for improvement.

I would like to sincerely thank the Evaluation Unit (WEQA) at SECO-WE and Syspons GmbH for their hard work and commitment. This report is more than just a collection of figures and ratings. It provides valuable insights that will help us to improve our projects and enhance their effectiveness.

The findings are encouraging. With 81 percent of evaluated projects achieving a satisfactory or highly satisfactory rating against OECD-DAC criteria, SECO-WE's projects are performing strongly.

Particularly noteworthy is the close alignment between internal and external assessments, which demonstrates that we are holding ourselves to high standards. Although external evaluations remain essential for credibility, internal tools such as Completion Notes are becoming increasingly important in capturing lessons learned and integrating them into future initiatives.

That said, there is always more that can be done. We will continue to strengthen the sustainability of our work, sharpen our understanding of how interventions lead to results, and pay closer attention to unintended effects that may influence outcomes. By applying the Theory of Change more consistently and examining causal links more deeply, we can better understand the impact of our actions. Expanding our pool of skilled evaluators will ensure we have access to the expertise we need to keep improving.

Monitoring and evaluation are core to our mandate. They make our work more effective, transparent, and ultimately more accountable to the people we serve. With new training and coaching programmes already underway, we are well positioned to build on these foundations.

This report comes at a timely moment. In line with the recommendations of the Council of States' Control Committee on Switzerland's international cooperation evaluation system, it provides renewed momentum to strengthen our evaluation practice. The key message is clear: quality matters more than quantity. By focusing on sound, meaningful causal analysis, we can best demonstrate the value and impact of our work.

I welcome the constructive recommendations set out in this report. Many have already been implemented, and we are pursuing the remaining ones with determination. This evaluation is a step forward, not an endpoint — one that we will continue to take with our partners and staff.

I hope that you will find this report both informative and inspiring.

Philipp Orga
Head of Operations
Economic Cooperation and Development Division
State Secretariat for Economic Affairs

TABLE OF CONTENTS

DAC	Development Assistance Committee
IC	International Cooperation
IMF	International Monetary Fund
OECD	Organisation for Economic Co-operation and Development
CC-S (GPK)	Control Committee of the Council of States (CC-S) (Geschäftsprüfungskommission des Ständerates)
PHRD	Peace and Human Rights Division
REQA	Rapid Evidence Quality Assessment
SDC	Swiss Agency for Development and Cooperation
SECO	State Secretariat of Economic Affairs
ToC	Theory of Change
ToR	Terms of Reference
WE	Economic Cooperation and Development division of SECO
WEOP	Economic Cooperation and Development: Operations
WEHU	Trade Promotion Section at SECO
WEIN	Infrastructure & Financing section at SECO
WEIF	Private Sector Development section at SECO
WEMU	Macroeconomic Support section at SECO
WEQA	Quality and Resources section at SECO-WE

EXECUTIVE SUMMARY

Performance & Quality Report 2023-24

Executive Summary

This meta-evaluation looks at how well SECO Economic Cooperation and Development projects are **performing** and how strong the **evidence** is that supports those assessments. The goal is not only to check results but also to see how evaluations themselves can be improved, so that SECO-WE becomes an even stronger **learning organisation**.

The review is based on **48 external evaluations**, combined with 60 internal completion notes, interviews, data analysis, and short case studies. Three approaches were used:

- a **Rapid Evidence Quality Assessment (REQA)** to test the reliability of evaluation evidence,
- an analysis of project **performance against OECD-DAC criteria**, and
- a **thematic content analysis** of how evaluations use Theories of Change (ToCs) and logframes, and what evidence evaluations provide for SECO-WE Business Line Impact Hypotheses.

Key Findings

- **Evaluation quality is mixed.** While methods are often adequate, many evaluations lack transparency, data quality, causal analysis, and proper triangulation of data. This weakens the strength of the evidence presented.
- **Project performance is satisfactory overall**, especially on relevance, coherence, and impact. However, sustainability is relatively weak: 41% of projects were rated unsatisfactory. While Effectiveness is rated mostly satisfactory, differential or unintended results are rarely considered.
- **Theories of Change (ToCs) are seldom used in evaluations. Also, evaluators regularly criticize logframes.** While ToCs are not mandatory at SECO-WE, they are frequently referred to in evaluations. However, even when mentioned, they are often missing, vague, or not assessed critically. Logframes are more commonly discussed in evaluations but often found to include unclear logic and unsuitable indicators.
- **Causal assumptions, including those suggested by SECO-WE Business Lines, are rarely tested.** Where examined, projects aligned well with Business Line Impact Hypotheses and underlying causal assumptions are more often confirmed than contradicted. However, most evaluations focus on activities and outputs, rather than outcomes and impacts, limiting learning on what works, why, and under what conditions.

What This Means

SECO-WE projects are generally on the right track, but both **implementation and evaluation practices** need strengthening. Evaluations must dig deeper into **sustainability**, **impact**, and **causality**. Likewise, better **results frameworks** (ToCs and logframes) could make projects easier to manage and allow for more insightful evaluations.

Recommendations

1. **Commission more and better causal evaluations** to understand what really drives results (High priority).
2. **Improve transparency in evaluation reports**, including more transparency on methods and data limitations (High priority).
3. **Strengthen M&E capacity** of programme managers and ensure projects budget for quality monitoring (Medium priority).
4. **Explore unintended effects** of projects through portfolio-level studies (Low priority).
5. **Reassess the added value of external DAC ratings** compared with internal completion notes (Low priority).

SECO-WE's projects are **mostly relevant, effective, and impactful**, with **sustainability assessed comparatively less strong than the other criteria**. Mixed **evaluation quality** and partially **modest results frameworks** limit both credibility of the evaluation and learning from it. Addressing these issues will make evaluations more useful for learning, steering, and accountability. This assessment primarily examined evaluations commissioned before SECO-WE implemented measures to improve evaluation quality, based on the recommendations of the Control Committee of the Council of States.

MAIN REPORT

Performance & Quality Report 2023 - 2024

1 Introduction

This assessment aims to assess the evidence quality and synthesise contents of project evaluations commissioned by SECO's Economic Cooperation and Development Division (SECO-WE) and its partners, as well as assess project performance based on these evaluations. The overarching objective is to generate insights that support **learning and steering** by identifying concrete ways to strengthen evaluation practice, evidence use and project cycle management within SECO-WE's operational sections. By meta-analysing all external evaluations and internal reviews, the report also ensures accountability for the assessed projects.

This report is one of the key moments where SECO-WE's Quality and Resources Section (WEQA) meta-evaluates and synthesizes evaluations to draw wider conclusions for the organisation. The report was expanded because of increasing demands on SECO-WE's evaluation system (see section 2). While this report continues to use the DAC rating methodology, it introduces the newly developed Rapid Evidence Quality Assessment (REQA) and employs thematic analysis on the use and quality of Theories of Change (ToCs) and logframes in evaluation.

The evaluation is guided by the following questions:

1. What is the quality of evidence provided in external project and independent evaluations? Which areas require improvement?
2. How do externally and internally evaluated projects perform against the OECD DAC evaluation criteria, and what explains variation in performance?
3. What insights do evaluations provide regarding the use and quality of Theories of Change and logframes? To what extent do these elements influence the effectiveness of evaluated projects?
4. What evidence do evaluations provide in relation to SECO-WE's impact hypotheses? Where are the strengths and gaps?

By increasing transparency on evidence quality in addition to project performance, this report reflects SECO-WE's commitment to being a learning organisation, continuously improving project management, as well as the quality of evaluation and enhancing the use of evidence in decision-making. It is written under the sole responsibility of the authors at WEQA and Syspons.

2 The Evaluation System's Operating Environment

Developments Shaping the Evaluation System in 2023-2024

In 2023 and 2024, several key developments significantly influenced the evolution of the evaluation system within Swiss International Cooperation (IC), setting the direction for the coming years.

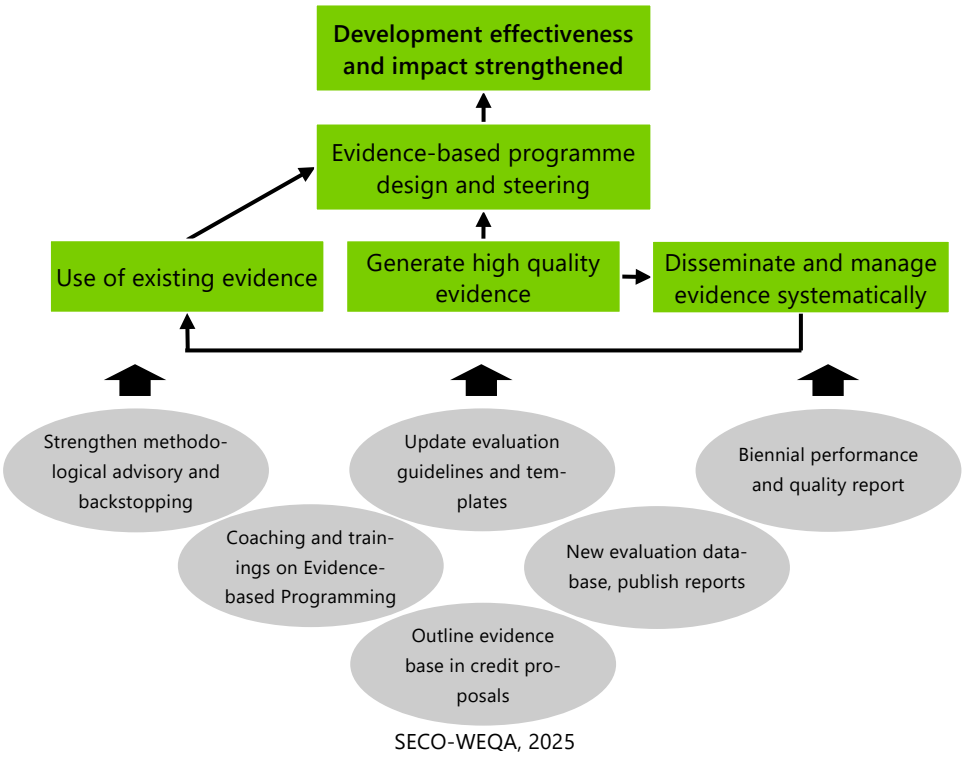
First, the **new IC Strategy 2025–2028**[1], with a financial envelope of CHF 11.12 billion, was developed and approved. While largely a continuation of the previous strategy, it provided an opportunity to refine results measurement frameworks where needed. In the area of Monitoring and Evaluation (M&E), SECO's Economic Cooperation and Development Division (SECO-WE) fine-tuned its eleven impact hypotheses[2] and updated its fifteen standard indicators.

Second, the Control Committee of the Council of States (CC-S) commissioned an **external evaluation of the Swiss IC's evaluation system**. The resulting report, published in November 2023, set out six key recommendations.[3] These included the need to strengthen collaboration among the three entities responsible for implementing the IC Strategy – namely, the Peace and Human Rights Division (PHRD), the Swiss Agency for Development and Cooperation (SDC), and SECO-WE. The evaluation underscored the importance of improving evaluation methodologies, better formulating and following up on recommendations, and ensuring alignment with overarching strategic objectives. It also called for increased transparency through improved publication practices and clearer accountability for results, thereby enhancing the credibility and utility of the evaluation function. In response to these findings and growing demands from operational sections, external stakeholders, Parliament, and the broader public, SECO-WE's Quality and Resources Section (WEQA) developed an Action Plan 2024-2026 to further strengthen its evaluation system.

Third, global geopolitical dynamics have contributed to an increasingly **complex and fragile operating environment** for international cooperation. Existing pressures - such as protracted crises, displacement and limited structural transformation in governance, service delivery, and economic systems – persist.[4] Against this backdrop, the relevance of Swiss IC remains undiminished. Public support remains strong: 86% of the Swiss population support either an increase (54%) or maintenance (32%) of current IC funding levels, while funding remains under pressure.[5]

[1] Federal Ministry of Foreign Affairs (FDFA). *International Cooperation (IC) Strategy 2025–28*.
[2] Thereof eight are thematic and three transversal impact hypotheses. The hypotheses are formulated as IF, THEN, BECAUSE statements.
[3] GPK-S (2023). *Wirksamkeitsmessung in der internationalen Zusammenarbeit. Bericht der Geschäftsprüfungskommission des Ständerates.*
[4] OECD (2025), *States of Fragility 2025*, OECD Publishing, Paris, https://doi.org/10.1787/81982370-en.
[5] ETH NADEL (2025). Swiss Panel Global Cooperation 2024. Swiss Attitudes Towards Migration and Foreign Aid.

Figure 1 | Evidence-based programming Theory of Change



SECO-WEQA, 2025

The grey shapes in figure 1 show key actions. The green shapes display how actions will lead to evidence generation, management and use. This will in turn contribute to considering evidence more systematically in project design and steering thus strengthening development effectiveness at SECO-WE.

Outlook

Having outlined the factors that shaped the system in recent years, what implications do this hold for the future?

First, work is underway to **digitalise** project monitoring. The impact hypotheses will be increasingly used to map existing evidence and will be updated as new findings emerge. This report constitutes an initial step in understanding how evaluations can help assess the plausibility of these hypotheses (see section 4: Evidence base of Impact Hypotheses).

Second, implementation of the **Action Plan 2024-2026** will remain a key priority. While SECO-WE aims to achieve 80% of the plan's objectives by 2025, the goal is to fully complete its implementation in 2026 and be in a position to demonstrate measurable change across the key areas illustrated in the green shapes (see figure 1). A full assessment of progress will be presented in the next Biennial Performance Report (2025-2026). Notably, a new Rapid Evidence Review (RER) service will be piloted in 2025, enabling operational units within SECO-WE to commission targeted evidence syntheses to inform their programming and strategic choices.

Third, while the global context is expected to become even more challenging, **spending for development cooperation globally is projected to decline**. In several countries – most notably in the United States – development budgets are being reduced significantly. In Switzerland, Parliament has lowered the 2025 allocation for international cooperation by CHF 110 million.

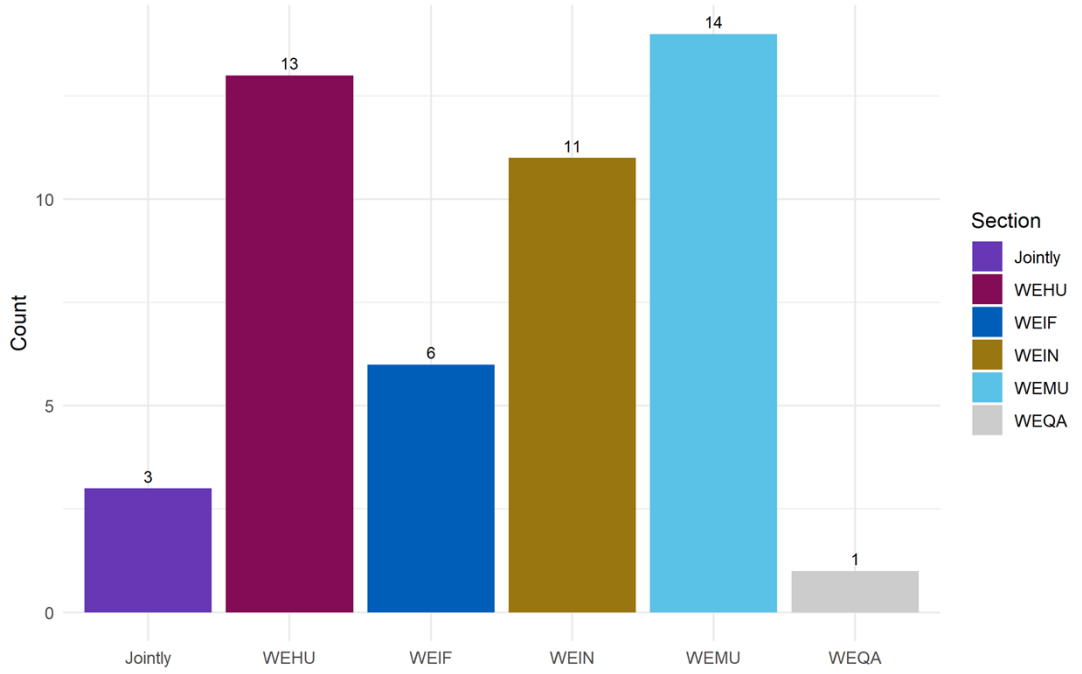
In this environment, it is more important than ever for Swiss IC to demonstrate the impact of its work. This cannot be done without robust evidence and credible evaluations. This report seeks to support that goal by documenting

recent improvements in the evaluation system and identifying patterns in what has worked, what has not, and why – thus providing strategic learning at the organisational level.

3 Approach and Methodology

The approach combines three complementary components: an evidence quality assessment (meta-evaluation), an assessment of project performance (meta-analysis of evaluation reports), and a thematic synthesis of content from evaluation reports; all complemented by short case studies. It is based primarily on 48 external evaluation reports[6], supplemented by 60 internal completion notes, case studies, and interviews. The meta-evaluation applies the REQA framework to assess evidence quality, while the meta-analysis uses the OECD Development Assistance Committee (DAC) criteria rating methodology to evaluate project performance, both developed by SECO-WE and SDC. Thematic synthesis is conducted through qualitative content analysis. A range of analytical techniques – including descriptive statistics, correlation analysis, and case studies – are employed to identify and understand patterns and derive conclusions. Most of the analysis is limited to what is documented in the final evaluation reports, complemented by purposefully selected case studies which add an additional layer of depth.

Figure 1 Evaluation reports by section (2023 and 2024)



[6] One evaluation, the "Independent evaluation of SECO-WE's climate approach", was not included in the OECD-DAC but only the REQA assessment.

Meta-evaluation using REQA

SECO-WE and SDC, with support of Syspons GmbH, have jointly developed the Rapid Evidence Quality Assessment (REQA) to **measure and strengthen the methodological robustness of evaluations in accordance with SEVAL and OECD standards**. REQA provides a systematic and reliable, yet rapid assessment of the quality of evaluative evidence in evaluation reports submitted to SECO-WE and SDC. It is applied retrospectively on all external project and independent evaluation reports submitted to SECO-WE and SDC (see Annex 4 for details).

By **making evidence quality visible**, REQA aims to establish a minimum standard for presenting evaluation methodology, raise awareness about methodological quality, and to enhance the usefulness of recommendations. By identifying recurring weaknesses in evaluation design, data collection, analysis, or recommendation formulation, REQA facilitates targeted improvements in evaluation practice and supports the broader goal of fostering evidence-based programming.

REQA draws conceptually on frameworks that assess the inherent evidence value of evaluation reports[7]. This focus allows REQA to **go beyond structural compliance with report formats and directly assess the credibility and accuracy of evaluation findings**. In doing so, it complements but does not replace broader quality frameworks such as the SEVAL evaluation standards[8] or OECD quality standards[9]. Notably, REQA covers approximately 37% of the SEVAL standards and 16% of OECD standards - those most relevant to methodology and recommendations.

REQA is more than a quality control instrument; it is a **strategic tool for learning and accountability**. It supports evidence use by making the strengths and limitations of evaluations more visible to decision-makers. At the same time, it signals to evaluators what the Swiss development cooperation considers as a high-quality evaluation. Over time, annual REQA analyses will help track progress in evaluation quality, build confidence in findings, and ensure that Swiss IC continues to meet high standards of credibility and transparency.

Scoring system and analysis

REQA is a rapid assessment tool built on a structured, rubric-based framework. It assesses evaluation reports along six core criteria[10]:

1. Adequacy of the evaluation for its intended purpose,
2. Transparency of methods,
3. Triangulation of data and perspectives,
4. Appropriateness of methods,
5. Strength of causal analysis (including attribution or contribution reasoning), and
6. Usability of recommendations.

The overall rating for each criterion is based on a weighted average across sub-criteria. The weighting was predefined and chosen such that more important sub-criteria have a stronger influence on the overall score for each criterion. The focus of the analysis is on the distribution of ratings and the average value of the scores[11]. Additionally, an exploratory analysis was conducted comparing the distribution and average value of the scores across different groups and correlation between different scores to identify patterns in the data.

---

[7] For example: Bond. (2018). *Evidence Principles*; DFID. (2014). *How to note: Assessing the strength of Evidence*; Apgar, M., Aston, T., Snijder, M., & Zwollo, T. (2024). Raising the Bar: Improving How to Assess Evidence Quality in Evaluating Systems-Change Efforts. Foundation Review, 16(2). doi:10.9707/1944-5660.1712

[8] SEVAL. (2016). *Evaluationsstandards der Schweizerischen Evaluationsgesellschaft*.

[9] OECD. (2010). *Quality Standards for Development Evaluation.* Paris: OECD Publishing. doi:10.1787/19900988

[10] While Criteria 2 to 6 are structured into sub-criteria, Criterion 1 is assessed as a single dimension. Each (sub-)criterion is evaluated on an unbalanced scoring scale ranging from 0 (not considered), 1 (inadequate quality), 2 (basic quality), 3 (good quality) to 4 (high quality), thus allowing for a nuanced yet efficient quality assessment. The only exception are criteria 5.1 and 5.2. about causal claims. They can also be rated "not applicable" if no causal claim is assessed in the evaluation. The use of rubrics promotes clarity, consistency, and transparency, while reducing subjectivity in the scoring process (the detailed scoring system and the process are presented in Annex 4).

[11] The average scores are calculated as unweighted means across evaluations. In the case of criteria 5.1 and 5.2 for causal claims, the average is calculated for those evaluations for which these criteria were assessed.

DAC Criteria Rating

To assess project performance, the meta-analysis applied the OECD Development Assistance Committee (DAC) evaluation criteria[12], using a structured rating methodology developed by SECO-WE and SDC. This standardised approach enables consistent, transparent, and quantifiable assessments of project results across six DAC criteria: relevance, coherence, effectiveness, efficiency, impact, and sustainability.

Scoring system

Each of the six DAC criteria is broken down into sub-criteria, typically three to five per criterion. These sub-criteria cover essential aspects of performance such as responsiveness to needs, quality of intervention design, operational efficiency, achievement of intended outcomes, and prospects for sustainability. For each sub-criterion, evaluators assign a score on a balanced four-point scale from 4 (highly unsatisfactory), 3 (unsatisfactory), 2 (satisfactory) to 1 (highly satisfactory). In case there is not sufficient information available for the assessment, this is also indicated for each sub-criterion (0=not assessed). For an overview of criteria and sub-criteria and details on the assessment process refer to Annex 5.

Analysis

The overall rating for each criterion is based on an unweighted mean across those sub-criteria which were rated, disregarding the sub-criteria rated as "not assessed". In case none of the sub-criteria was rated, i.e. each was marked "not assessed", the overall rating is "not assessed" too. The analysis focuses on descriptive statistics at the level of individual DAC criteria (e.g. average scores, distribution patterns). The data are not aggregated into a single overall performance score per project. Additionally, an exploratory analysis was conducted comparing the distribution and average value of the scores across different groups and correlation between different scores to identify patterns in the data.

It is important to note that not all evaluations apply all DAC criteria[13], as some may be excluded based on the Terms of Reference or the project's evaluability. In fact, while at least one sub-criterion of Effectiveness is rated for about 90% of evaluated projects, Impact was assessed for only 2 in 3 evaluated projects. Moreover, external and internal evaluations cover different projects and cannot be compared directly. Finally, the evaluation timing is different across projects, with 47% being mid-term, 42% end-term, and 7% ex-post evaluations[14], such that the performance refers to performance at different points in time. While the methodology enables aggregation at the level of individual DAC criteria, findings are not generalisable to the full SECO-WE portfolio due to the non-representative sample of evaluated projects.

[12] OECD (2019), *Better Criteria for Better Evaluation: Revised Evaluation Criteria Definitions and Principles for Use*, OECD Publishing, Paris, https://doi.org/10.1787/15a9c26b-en.

[13] OECD (2021), *Applying Evaluation Criteria Thoughtfully*, OECD Publishing, Paris, https://doi.org/10.1787/543e84ed-en.

[14] The remaining 4% were differently categorized.

Thematic synthesis of content from evaluation reports

The content synthesis aims to consolidate thematic and methodological insights from SECO-WE's 47 project evaluation reports to assess how projects and evaluations articulate and apply results logics, specifically Theories of Change and logframes, and to examine their alignment with Business Line Impact Hypotheses[15]. In doing so, the synthesis provides both an **overview of portfolio content and reflections on the evaluative treatment of results logics**.

The synthesis was based on a qualitative content analysis of 47 evaluation reports, conducted using the software MAXQDA. The reports covered project evaluations completed in 2023 (25 reports) and 2024 (22 reports) and spanned different thematic areas and organisational units. The three Business Lines with the largest coverage in terms of evaluation reports were selected for an in-depth analysis of causal relationships (Business Lines 1.1, 2.2 and 3.1)[16], together accounting for 60% of the evaluation reports, while the remaining 17 reports related to any of the other five Business Lines.

The analysis followed a **multi-step approach combining deductive structure and inductive refinement**. A coding system was developed based on the central evaluation questions, covering the quality of Theories of Change and logframes, Business Line-specific content (e.g. contextual factors, causal claims, alignment with Business Line logics), and accompanied by detailed guidance to ensure consistency. Evaluation reports were imported into MAXQDA, enriched with background data (e.g. Business Line affiliation, REQA and DAC scores), and coded in two iterations: an initial deductive round using the predefined structure, followed by an inductive clustering and refinement of subcodes. Memos supported interpretation throughout. Finally, the data was analysed both qualitatively and quantitatively, using frequency counts, in-depth content analysis, and intersections with REQA and DAC scores to explore how evaluation content related to methodological quality or evaluative focus (detailed process description presented in Annex 6).

Case studies and triangulation

The analyses were further supplemented by case studies, i.e. selected evaluations, whose background and follow-up were further explored and supplemented by qualitative interviews. The selection of the **six cases** followed a **purposive sampling approach aimed at ensuring variation along key dimensions of interest**. Specifically, the sample sought to include evaluations characterized by both high and low evidence quality as measured by the REQA framework, as well as varying levels of project performance, specifically effectiveness based on the OECD-DAC criteria. In addition, efforts were made to ensure representation from at least three different sections. Beyond project evaluations, the portfolio-level independent climate evaluation was included in the sample as well. Evaluations were also selected based on specific knowledge interests or because they appeared to reflect particularly good practices worthy of deeper examination[17]. DAC ratings and REQA scores for the selected evaluations can be found in Annex 6. In total, **12 interviews** were conducted with the respective SECO project managers, external evaluator(s), and responsible person(s) on the partner side (where applicable, i.e. in cases where the evaluation was not commissioned by SECO). These contextualized and illustrated the findings from the three assessments, whereby particular attention was paid to the context and motivation behind the evaluation, participation, and the perceived usefulness. The interviews were conducted in a semi-structured manner using an interview guide and documented through summary notes, complemented by full transcripts. The results of these case studies are presented in dedicated boxes throughout the report, offering insights into recurring, strategically relevant topics identified through the interviews and case evidence.

[15] SECO-WE has operationalized the IC strategy through eight sectoral and three cross-cutting Business Lines. Each Business Line is guided by an Impact Hypothesis articulated in the form of an "If…then…because" statement.
[16] 15 reports related to Business Line 1.1 Growth-promoting policy, four to Business Line 2.2 Integration into Value Chains, and ten to Business Line 3.1 Urban Development.
[17] For example, emerging from the content analysis of the use of Theories of Change and logframes.

Limitations

The present analysis is subject to several limitations that should be considered when interpreting the findings:

- **Scope of the review:** The assessment focused exclusively on evaluation reports. Supporting documents such as Terms of Reference (ToR), inception reports, or other preparatory materials were not considered. As a result, important contextual or methodological elements that may have been documented elsewhere remain outside the scope of this assessment. Moreover, information not included in the final report could not be assessed, following the principle that "what is not written cannot be evaluated." Similarly, the use of Theories of Change or logframes in evaluations (content analysis) could only be meaningfully assessed when such frameworks were explicitly included or at least discussed in the evaluation report. Contextual information was, however, collected in the context of the five in-depth case studies, which were used to complement and illustrate the findings of the broader review.
- **Emphasis on methodological quality of evaluations:** The REQA framework, which was developed for and applied throughout this assessment, focuses primarily on methodological rigor. Aspects such as participation and usefulness of the evaluation - while important - were only considered for the case studies (interviews). In addition, causal analysis was relatively rare across the reports (22 of 48 reports), which limits the assessment of causal claims and the strength of causal evidence presented in chapter 4.
- **Variations in applying DAC criteria:** Not all evaluations assessed all OECD-DAC criteria, as the decision to include specific criteria is often guided by learning interests or steering needs. Moreover, the DAC ratings are based on varying consideration of sub-criteria across projects. This means that the same overall DAC criterion may reflect different aspects in different evaluations, limiting comparability. Therefore. the discussion of the different DAC criteria always includes a discussion of the different sub-criteria. Further at the end of the section there is a discussion on the extent to which the different criteria were assessed across projects.
- **Representativeness and generalizability:** Importantly, it should be noted that the evaluated projects do not constitute a representative sample of the SECO-WE portfolio. Evaluations at SECO-WE are commissioned in response to specific learning interests or steering needs identified by operational sections (and not by random selection). This approach is considered essential to enable evaluation to realise its full potential. Consequently, no generalizable conclusions can be drawn about the portfolio as a whole. This also applies to the thematic content analysis, which focuses on three out of eight impact hypotheses only, and remains restricted to those evaluations which actually provide information on causal claims – whose number is limited.
- **Different rating cycles:** The DAC assessments of external evaluations conducted in 2023 and 2024 were carried out by different contractors. As a result, minor inconsistencies in interpretation and scoring cannot be ruled out. The REQA assessment of external evaluations conducted in 2023 and 2024, in contrast, was implemented by one contractor in 2025.

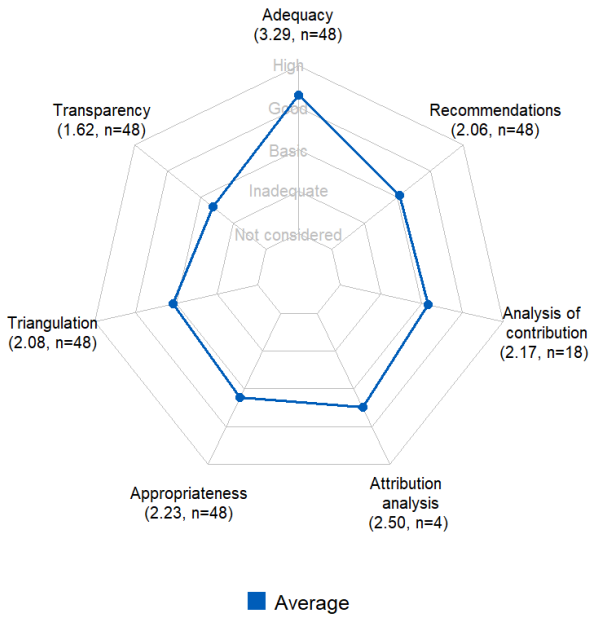
These limitations underscore the importance of interpreting the results of this evaluation as indicative rather than definitive, and as a contribution to learning and quality improvement rather than a comprehensive portfolio-wide assessment.

4 Findings

The following chapters present analysis results from the evidence quality assessment (meta-evaluation), the assessment of project performance (meta-analysis), and the thematic synthesis of content from evaluation reports on the use of Theories of Change and logframes in evaluations, as well as on evidence in relation to SECO-WE's impact hypotheses.

Evidence Quality of Evaluations (REQA)

Figure 2 | Average REQA Scores



For most dimensions of the REQA there is ample room for improvement across evaluations (see Figure 2). With the exception of *Adequacy*, the average REQA score is just above, or in the case of *Transparency* just below 2, corresponding to **basic quality**. For each of these criteria, only very few evaluations (2-5%) can be considered as very high quality. More than half of the evaluations scored basic (score equivalent to 2), inadequate (1), or did not consider the criterion in the report (0).

The analysis reveals a mostly positive correlation between the scores, indicating that while each criterion captures distinct aspects of evidence quality, they are interrelated. In general terms, high quality in one area is generally associated with higher quality in the other. There is a strong tendency for reports receiving a high rating for *Transparency* to also receive higher ratings for *Triangulation* and *Appropriateness*; and similarly for *Appropriateness* and *Triangulation*.[18] For an overview of correlation coefficients refer to Annex 11.

A comparative analysis of REQA scores between 2023 and 2024 reveals an overall positive shift in evaluation quality. Most score dimensions showed an increase in their average values across evaluations, with the most notable improvements observed in *Transparency* and *Appropriateness*, where the median scores rose by one full level. For *Adequacy* the average score showed a negligible decrease compared to the increases for other dimensions (see

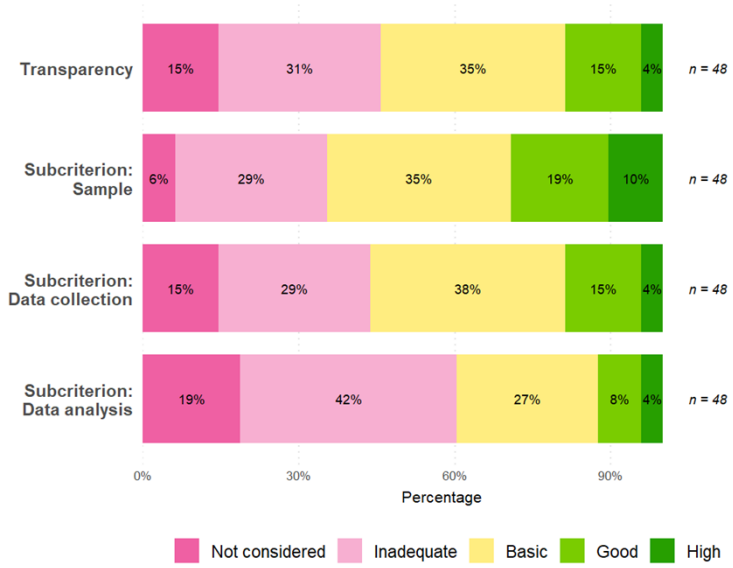
[18] Note that there are only four ratings for the evidence quality of attribution claims such that these numbers are not particularly insightful and driven by single reports.

in Annex 11). To explore whether these differences reflect practically meaningful changes, non-parametric statistical tests were applied. While these tests suggest that the observed shifts may not reflect consistent or statistically meaningful patterns, they nonetheless support a cautiously optimistic interpretation of incremental improvements in evaluation quality.[19]

Adequacy

In 92% of cases, an evaluation was found to be an adequate tool for its stated objectives. This reflects a general assessment based on the stated objectives and purpose rather than considering the specific evaluation approach, design, or methods which is covered under the criterion for Appropriateness. Among half of these, the use of an evaluation was further well justified in the report. In 4% of cases, it is unclear whether another tool would have been more adequate and in 4% there was no purpose stated, or an evaluation was assessed an inappropriate tool (for instance, where the report's focus was primarily on exploring options for a subsequent project phase).

Figure 3 | Transparency criteria



Note: Percentages may not total 100% due to rounding.

Transparency

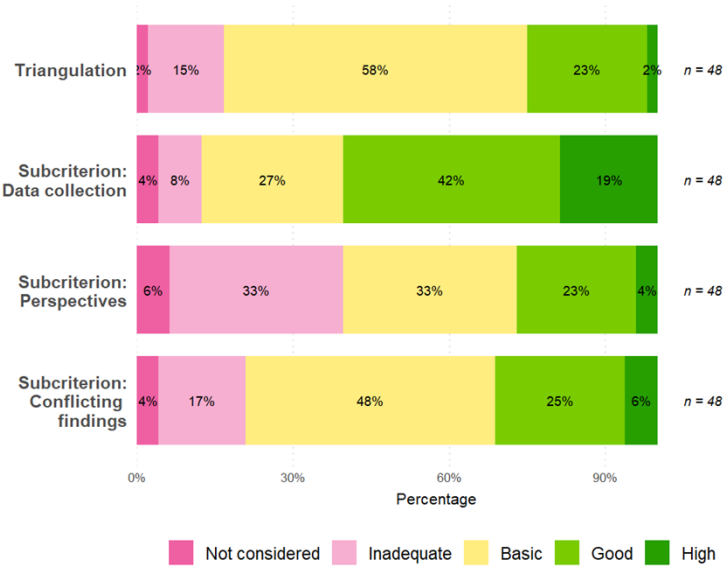
The *Transparency* sub-criteria are rated rather poorly. This criterion assesses whether the evaluation clearly describes and justifies the i) sample, ii) data collection and iii) analysis methods. Sample descriptions are often minimal, analysis methods vague or omitted, and limitations regarding the sample, data collection or analysis are rarely discussed in a meaningful way. When limitations are mentioned, they usually focus on logistics without a reflection for what it implies for data quality, evidence strength or the analysis. The different sub-criteria are strongly correlated (0.7-0.8) implying that reports tend to be similarly transparent regarding sample, data collection and analysis.

[19] Mann-Whitney U tests for the equally of distributions between the two years could not be rejected at the 10% significance level. While all evaluation reports are assessed and an improvement in most REQA scores is observed between 2024 and 2023, statistical tests can be useful to distinguish strong from weak patterns.

Triangulation

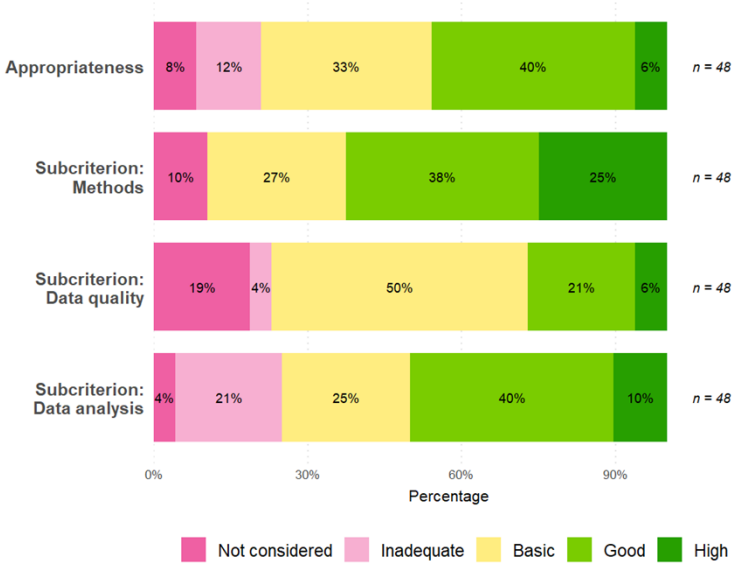
The *Triangulation* sub-criteria show mixed results with about half of the evaluations rated as basic overall. This criterion assesses the extent to which i) multiple data collection methods are used, ii) different stakeholder perspectives considered and iii) conflicting or nuanced findings are discussed. Many evaluation conduct interviews (94%) and other methods such as surveys (27%), focus group discussion (25%), or systematic or extensive document reviews (29%) are frequently employed while analysis of administrative data[20] (8%) is rather rarely done. While most evaluations (61%) use multiple data collection methods, they rarely explain why these methods were chosen to complement each other (19%). The origin of findings is often unclear, and multiple perspectives are seldom explored in detail. Conflicting results are almost never reported or discussed, though many reports still display a degree of nuanced analysis. Triangulation of methods or perspectives is rarely visible from the report. There is a moderate correlation between the sub-criteria, indicating that ratings across them vary, though with a similar tendency.

Figure 4 | Triangulation criteria



Note: Percentages may not total 100% due to rounding.

Figure 5 | Appropriateness criteria



Note: Percentages may not total 100% due to rounding.

Appropriateness

The *Appropriateness* sub-criteria show mixed results but almost half (46%) are rated good. This criterion assesses whether the evaluation's i) approach, design and methods, ii) data and sample, iii) and analysis are suitable and can generate reliable findings and conclusions. Less than one-third of the evaluations include a theory-based approach (29%) and most are based on an ex-post-facto (46%) or case-based (42%) design. Case-based designs typically are also ex-post-facto designs drawing on cases. Only in rare instances are (quasi-)experimental methods used (8%). Further, 6% of evaluations are explicitly utilisation-focused and 19% are described as participatory. The overall approaches and methods of evaluations are generally assessed as suitable (63%). Data quality, on the other hand, is often poor or cannot be assessed (23%) due to a lack of transparency. Reports rarely include an assessment of data quality or related limitations. Analyses are often good (40%) and there are also few exemplary cases (10%). However, frequently analyses are found to be inadequate (21%) or appear fully absent or severely lack clarity and coherence (4%). Findings are frequently not transparently substantiated by underlying data, and limitations are rarely acknowledged – both of which undermine the credibility and trustworthiness of the conclusions

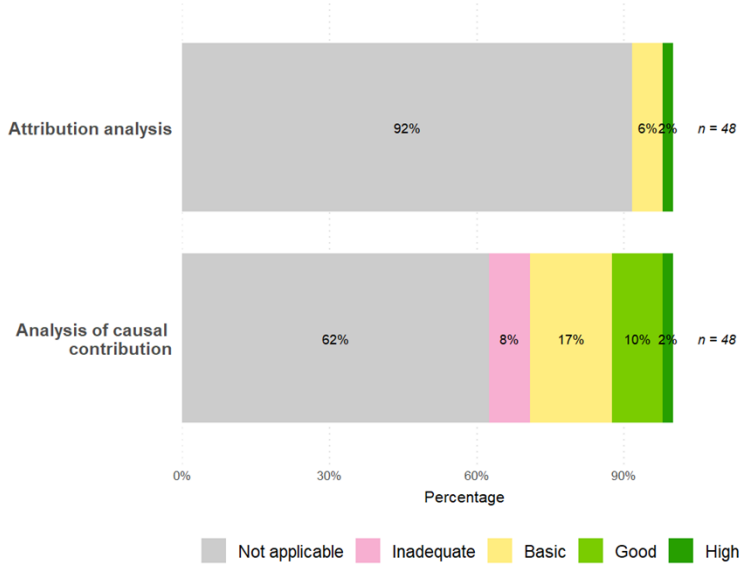
[20] This can include a wide range of data such as implementation data, including financials, but also records from civil registries.

presented. The sub-criteria are moderately correlated, indicating some consistency in how evaluations perform across sub-criteria.

Causal analysis

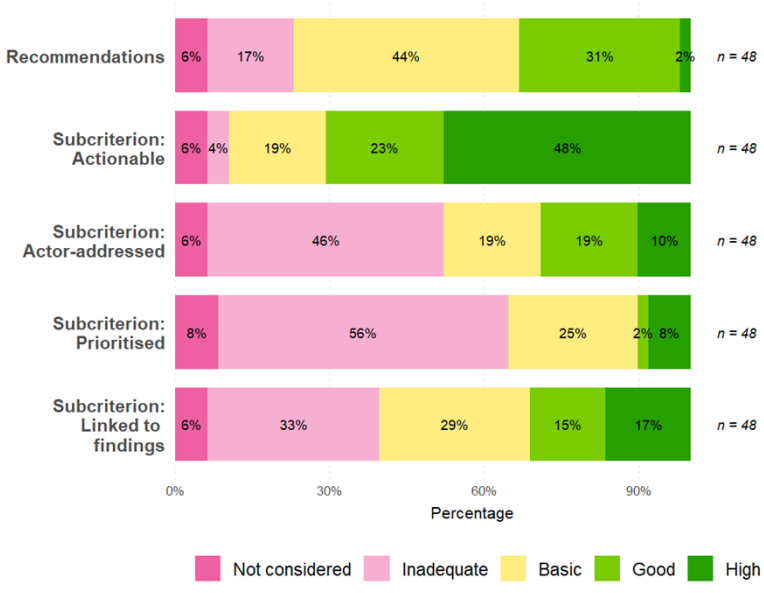
Most evaluations do not formulate or rigorously analyse explicit causal claims. The REQA assessed whether an evaluation identifies the intervention's role in producing observed changes based on a suitable analysis. Only four evaluations (8%) include an attribution analysis and 19 (38%) of the 48 evaluations reportedly conduct an analysis of contribution[21]. One of the attribution analyses is rated highly, while the remaining three are rated basic lacking a reflection on bias and underlying assumptions. The analyses of contribution are similarly mixed in quality: One is rated highly, but four are considered rather inadequate missing an analysis for the most part. The remaining 13 are rated as basic (8), often because alternative explanations are not considered but sometimes due to data limitations, or good (5).

Figure 6 | Causal analysis criteria



Note: Percentages may not total 100% due to rounding.

Figure 7 | Recommendations criteria



Note: Percentages may not total 100% due to rounding.

Recommendations

The quality of the recommendations provided by the evaluations leaves room for improvement across several sub-criteria, though on balance, most reports fall within a basic to good range overall. In contrast to the other criteria, this criterion does not capture the evidence value of the evaluation but assesses the extent to which recommendations in the evaluation are i) actionable, ii) actor addressed, iii) prioritised, and iv) linked to findings. While many recommendations are deemed actionable, they are rarely targeted at specific actors, typically addressing only the organisation or project in general terms. Prioritisation is almost entirely absent, despite a substantial number of recommendations being offered in many cases. The degree to which recommendations are clearly derived from and linked to findings varies significantly. There is a low correlation among sub-criteria which explains why the overall score is often basic or good despite frequent inadequate scores in sub-criteria.

[21] This does not necessarily refer to a contribution analysis as developed by John Mayne but, in principle, could cover other approaches, e.g. process tracing. Furthermore, it is based on how the evaluation report presents its evaluation approach and design, i.e. if it explicitly aims to address contribution claims.

INFO BOX 1: INTERNATIONAL COMPARISON - EVIDENCE QUALITY OF EVALUATIONS

A meta-evaluation conducted by the German Institute for Development Evaluation (DEval) mentions that 73% of evaluations barely or partly describe appropriateness and limitations of the applied methodology.[22] In a peer reviewed article, Stefan Silvestrini of the Centre for Evaluation (CEval) identifies significant weaknesses in methodology across ten meta-evaluations and 407 evaluation reports from agencies in different European countries.[23] In another peer reviewed article, Ellen Mastenbroek and colleagues conclude that the methodological quality across a sample of 216 ex-post evaluations on EU legislation by the European Commission is disappointing.[24]

These findings highlight that methodological weaknesses are not limited to SECO-WE. They also exist in numerous other European countries and across various policy areas, as also pointed out in the report by the Control Committee of the Council of States (CC-S) [25]. With REQA, SECO-WE and SDC have developed a framework to make these weaknesses visible in a nuanced way, which allows addressing them going forward.

[22] DEval (2022). *Meta-evaluation on the Quality of (Project) Evaluations in German Development Cooperation*, p. 42.
[23] Silvestrini, S. (2023). Meta im Quadrat - Was wir aus Meta-Evaluationen lernen können. Zeitschrift fur Evaluation, 22(1), 120-146. doi:10.31244/zfe.2023.01.07
[24] Mastenbroek, E., van Voorst, S., & Meuwese, A. (2016). Closing the regulatory cycle? A meta evaluation of ex-post legislative evaluations by the European Commission. *Journal of European Public Policy, 23*(9), 1329–1348. https://doi.org/10.1080/13501763.2015.1076874
[25] GPK-S (2023). *Wirksamkeitsmessung in der internationalen Zusammenarbeit. Bericht der Geschäftsprüfungskommission des Ständerates.*

CASE STUDY INSIGHT 1: PURPOSE AND SCOPE OF EVALUATIONS – BALANCING EXPECTATIONS AND PRACTICALITY

Case studies shed light on how expectations shape the purpose and scope of evaluations, highlighting the specific strengths of both rigorous designs and smaller-scale evaluations from a user perspective. A recurring theme across the case studies is the **importance of aligning the focus of evaluations with the specific needs of projects rather than evaluating all six OECD-DAC criteria**. The latter approach was seen as allowing for a more targeted evaluation that can yield more relevant and actionable insights.

As to the purposes, evaluations are sometimes referred to as **communication tools** to backup decisions and provide justification for the next steps. As noted by one SECO-WE programme manager, the evaluation "gave us backing", providing a well-founded case for moving forward into a further project phase. In this and another case, the evaluation was used as accountability tool, particularly when stakeholders wanted to show that the project was being conducted effectively and that future phases would be designed based on solid evidence.

Evaluations were also found to be crucial in **addressing and identifying challenges within the project**. In some cases, such as the 'External Evaluation – Swiss Trade Policy and Export Promotion Project Vietnam' (WEHU 227), evaluations helped pinpoint issues within the project that had been lingering (see also Insight 4). Evaluators were often seen as providing a valuable independent perspective, which was crucial for understanding and addressing these challenges. According to some interviewees, a participatory approach (see Insight 2) could provide practical and well-founded recommendations that could compensate for smaller-scale evaluations if the purpose is to provide 'food for thought' for programme design. For example, the 'Evaluation of the Colombia Más Competitiva Programme' (WEHU 226) helped identify that the project was too broad, with too many areas of focus, which made it difficult to achieve the intended objectives. As a result, the evaluators' recommendation to narrow the focus was seen as essential, particularly in light of budget cuts. This links to another common purpose of evaluations identified in the case studies, which is their role in **preparing for another phase of the project**. The majority of the cases (evaluations) was specifically designed to inform the design and eventually also generate funding of future phases.

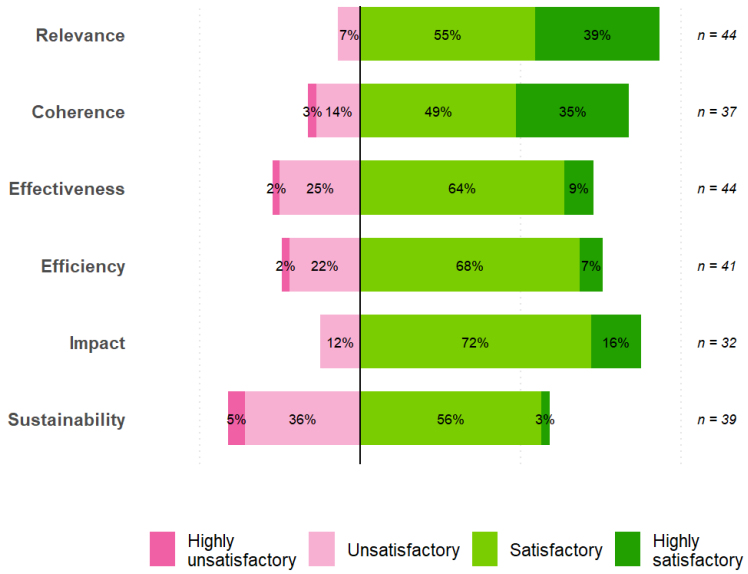
> "Both evaluations were useful. While one focused on specific questions they [the project team] had, the first was addressing other aspects. It depends on the situation: for WEHU 227, it was about answering urgent questions; the first evaluation, WEMU 121, was a larger exercise for a bigger project, where accountability was also important. In the case of WEHU 227, this in turn would have been too early. It's about having the right tool for the right situation." (SECO-WE project manager)

Finally, the scope and focus of evaluations are key factors in determining their usefulness. Case studies have shown that **evaluations should be flexible enough to allow for a particular emphasis on the most relevant issues for the project**. This can be achieved, for example, by partners providing input on the Terms of Reference or by evaluators exploring the potential and need for a specific focus during the inception phase. It was stated that evaluations which try to cover every aspect of a project or adhere too strictly to all OECD-DAC criteria can sometimes miss important nuances.

Project Performance along DAC Criteria

Overall, the evaluated projects perform mostly satisfactorily across OECD-DAC criteria. Especially Relevance, Coherence, and Impact are assessed as satisfactory for more than 85% of evaluated projects and among these, frequently highly satisfactory. While Effectiveness and Efficiency are also primarily assessed satisfactorily, unsatisfactory ratings are more frequent and highly satisfactory ratings remain below 10%. Overall, Sustainability is assessed the lowest with 41% of evaluated projects receiving an (highly) unsatisfactory assessment.

Figure 8 | Overview of OECD-DAC criteria



Note: Percentages may not total 100% due to rounding.

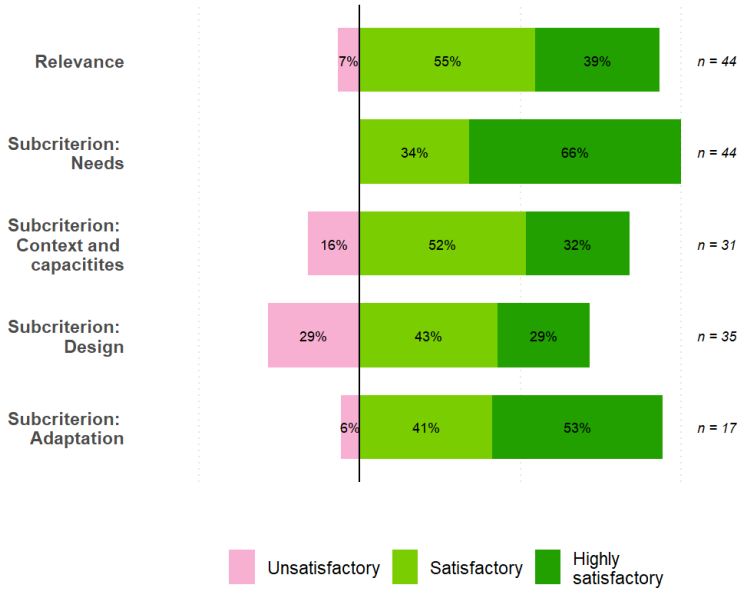
While the projects' performance is found to be mostly satisfactory, previous performance reports generally found higher shares of satisfactory projects. These changes should however not be interpreted as a decrease in performance of SECO-WE's portfolio of projects, since statistical uncertainty, selection bias and changes in assessment processes are likely contributors to observed changes. For more details on differences to previous years, refer to Annex 7.

The analysis further reveals that projects which are assessed well against one of the OECD-DAC criteria have a tendency to also be assessed well against others. Generally, there is a moderate positive correlation between the ratings of different OECD-DAC criteria. This is to be expected as the criteria measure different aspects of project performance which should be associated. For an overview of correlation coefficients refer to Annex 10.

Relevance

The projects are generally assessed as strong across the relevance sub-criteria. All evaluated projects, with few exceptions where no related information is available, demonstrate a satisfactory or highly satisfactory alignment with the needs of target groups. Projects further perform well overall regarding their alignment with contextual factors and the capacities of target groups. Design considerations are evaluated largely favourably, but deficiencies are more commonly reported, e.g., that projects are overly ambitious or lack an impact logic, i.e. it is not well articulated or captured in the projects' logframes or ToCs how the project results would lead to impacts. The projects' adaptability is typically assessed positively but only assessed in about one-third of cases. Often this

Figure 9 | Relevance criteria



Note: Percentages may not total 100% due to rounding.

adaptability does not correspond to changes in the projects' results logic but can include changes in modes of delivery or specific content of capacity development. A moderate to high correlation among sub-criteria suggests they capture distinct facets of relevance, yet project performance tends to follow a consistent trajectory across them.

Figure 10 | Coherence criteria

Note: Percentages may not total 100% due to rounding.

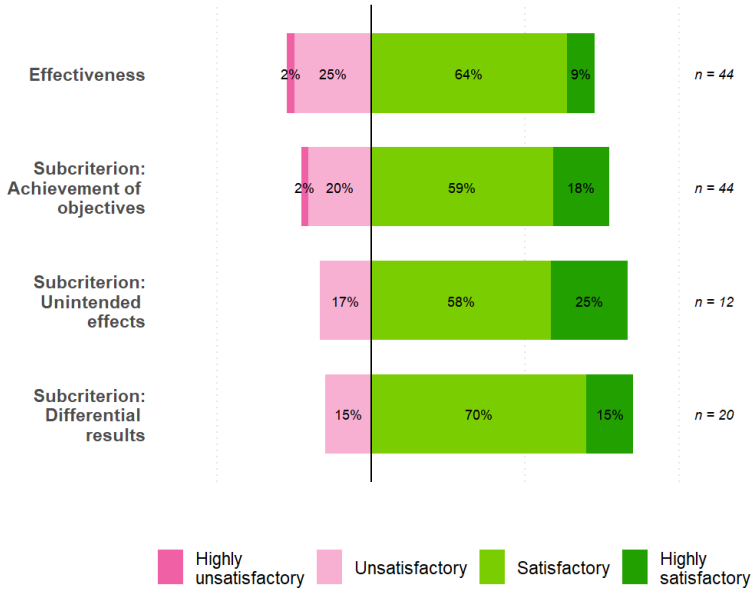
Coherence

The assessment of Coherence, across the respective sub-criteria, is generally positive. The assessment of policy alignment is often limited to a general affirmation that the intervention aligns with SECO's overarching goals or, less frequently, with specific strategies. The assessment is always (highly) satisfactory. Internal and external compatibility are more mixed. While most evaluated projects are rated as satisfactory, more than every fifth project receives an unsatisfactory, very few even a highly unsatisfactory assessment. Reasons behind unsatisfactory scores are typically due to a lack of synergies or coordination with other Swiss interventions. The compatibility assessments often refer to similar objectives or synergies across projects, but a deeper analysis is missing. Policy alignment and internal compatibility within the Swiss international cooperation system is addressed in less than half of the evaluations. Notably, policy alignment is less frequently addressed in evaluations commissioned by partners (about 1 in 3 cases) compared to evaluations commissioned by SECO-WE (about in 2 of 3 cases). Similarly, internal compatibility is addressed in 4 of 5 evaluations commissioned by SECO-WE but only in 2 of 3 cases when commissioned by the partner. Often these evaluations focus on the internal coherence within, for example, the respective implementing multilateral organisation.

Effectiveness

Effectiveness is generally rated satisfactory. While 59% of projects are assessed as satisfactory in terms of achieving their objectives, 18% demonstrate a high achievement; shortcomings are identified in 22%. Unintended effects are addressed in only every fourth report, though when they are, they tend to be positive (i.e. positive "side-effects" of projects) (83%). Differential results, i.e. the extent to which the intervention results were inclusive and equitable amongst beneficiary groups, are mostly assessed as (highly) satisfactory (85%). The moderately strong correlation among sub-criteria suggests consistency in overall effectiveness, despite the criteria capturing distinct dimensions.

Figure 11 | Effectiveness criteria



Note: Percentages may not total 100% due to rounding.

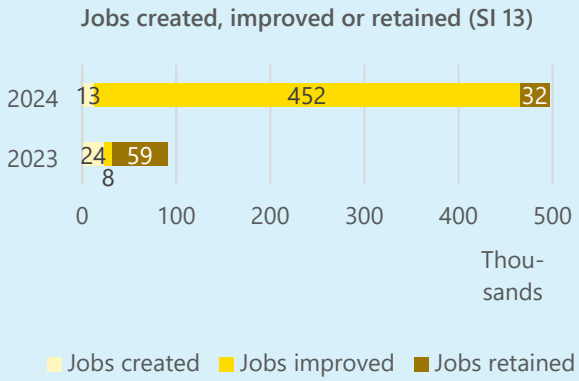
INFO BOX 2: STANDARD INDICATORS PROVIDING INSIGHT INTO PORTFOLIO EFFECTIVENESS

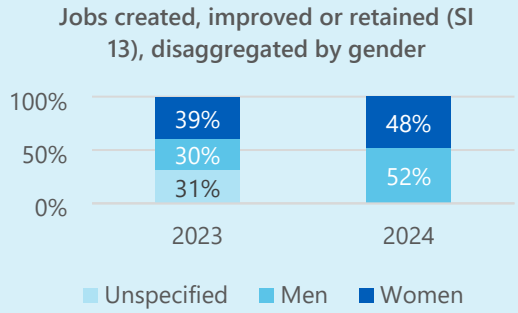
SECO-WE's performance measurement system includes a set of **15 standardized indicators (SI)**. This section presents the results for 2023 and 2024[26]. Annual fluctuations must be interpreted with utmost caution, because not all projects report results on a yearly basis.

In 2023, **capital mobilisation or enablement for development outcomes** totalled USD 2.97 billion, decreasing slightly to USD 2.9 billion in 2024 (SI 6). SECO-WE projects also contributed to the mobilisation of USD 1.84 billion in domestic resources in 2023, followed by USD 1.56 billion in 2024 (SI 2). A substantial share of these results was generated through the Swiss Investment Fund for Emerging Markets (SIFEM).

In the area of **climate change mitigation**, emissions reductions reached 10.6 million tonnes of CO_2 in 2024, up from 6.12 million tonnes in 2023. A significant share of this increase stems from SIFEM and the harmonisation of calculation methodologies. Energy savings and additional renewable energy generation amounted to 5.19 billion kWh in 2023 and 4 billion kWh in 2024, again largely driven by SIFEM-supported projects. SIFEM being a substantial driver of results in these areas reinforces SECO-WE's decision to conduct an independent evaluation of its impact and effectiveness. The climate change mitigation results indicate that climate mainstreaming efforts initiated following the independent climate evaluation and new climate policy in 2023 are yielding tangible outcomes.

Job creation, retention, and improvement rose markedly - from over 90'000 jobs in 2023 to 497'000 in 2024 (see graph to the right). This sharp increase is mainly due to the inclusion of over 450'000 improved jobs reported by the International Labour Organization's (ILO) Better Work programme, which did not report results in this area in 2023.

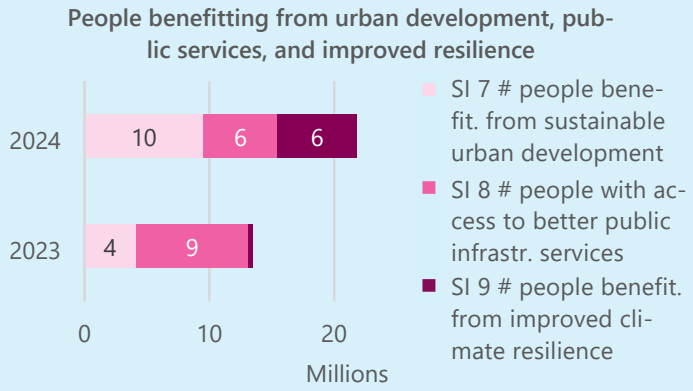




As shown on the graph to the left, gender-disaggregated data became more widely available in 2024, with jobs for women and men recorded at nearly equal levels.

Access to capital also improved: approximately 10'000 individuals gained access in 2023, compared to 20'000 in 2024 (SI 12). These results were primarily driven by the Women Banking Champions programme in the MENA region and were largely achieved by women-led enterprises. The number of companies gaining access to new markets or value chains (SI 15) declined from 3'150 in 2023 to 850 in 2024.

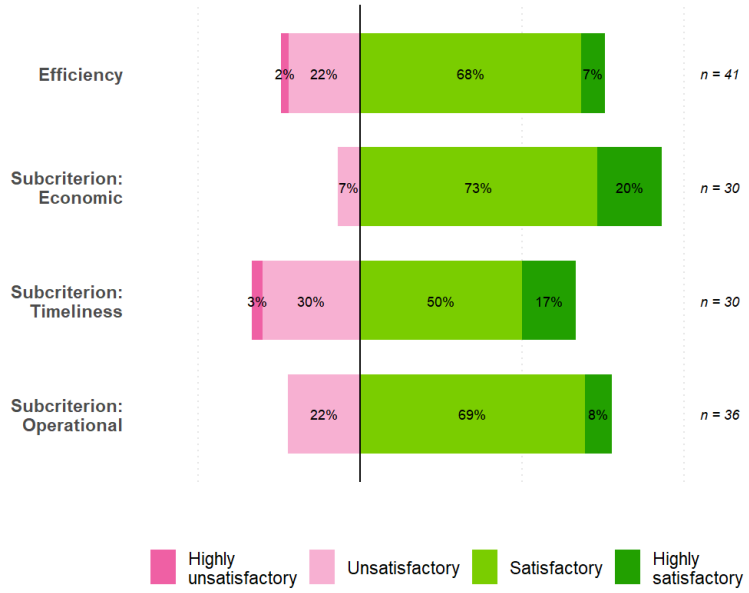
As the next graph shows, the number of **individuals who benefited** from SECO-WE infrastructure interventions increased from 13.48 million in 2023 to 21.82 million in 2024. The biggest contribution stems from the Sustainable Urbanization in Indonesia programme. In addition, 58'000 people benefited from measures to improve working conditions in 2023, compared to approximately 21'000 in 2024 (SI 14). Likewise, the number of people trained through SECO-WE projects dropped from 140'000 in 2023 to 72'000 in 2024 (SI 4).



[26] 2024 figures remain preliminary and may be adjusted as additional data become available

Efficiency

Figure 12 | Efficiency criteria



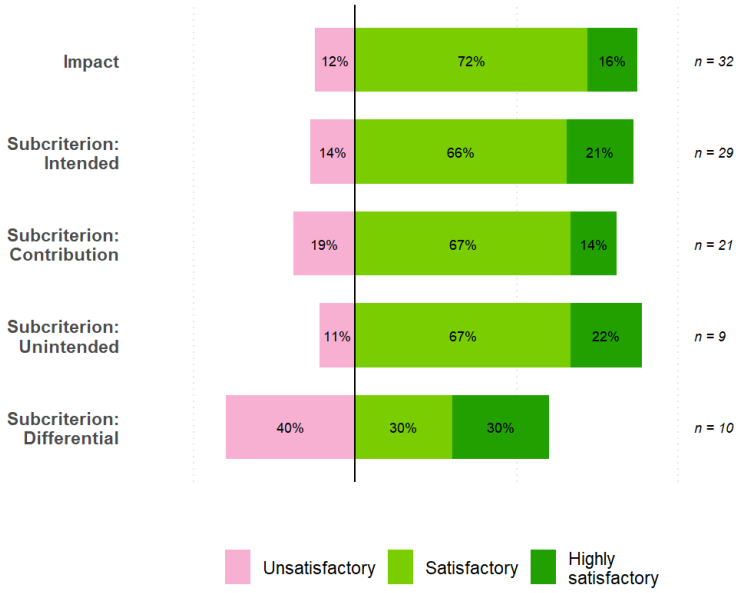
Note: Percentages may not total 100% due to rounding.

Efficiency is mostly rated satisfactory (68%) but with few highly satisfactory projects (7%) and 24% (highly) unsatisfactory projects. Economic efficiency is assessed as (highly) satisfactory in 93% of the cases; however, analyses are often based on weak data or remain superficial. Delays are common, but larger deficiencies are only discussed in 33% of evaluations and in most cases, delays are minor or well mitigated (50%). Operational efficiency tends to be assessed favourably, with low levels of identified shortcomings in most reports (69%), though notable issues are reported in roughly 22% of cases, stemming e.g. from unclear roles and staff turnover, the lack of a steering committee or overly complex project structures. Correlations suggest moderate alignment between economic efficiency and the other sub-criteria, but no relationship between timeliness and operational efficiency.

Impact

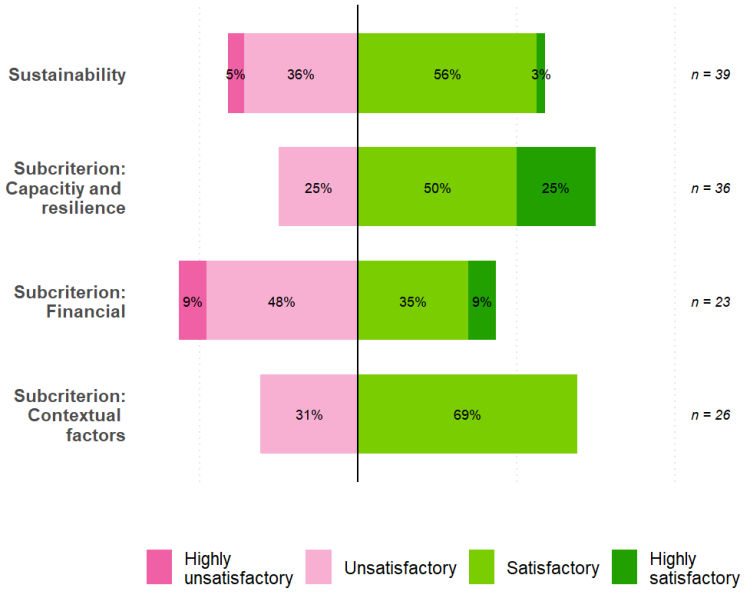
The (expected) Impact of SECO-WE's projects is generally rated as satisfactory. For each sub-criteria of intended impacts, contribution to intended impacts and unintended impacts about 2 in 3 projects are assessed to be satisfactory. While some (11%-19%) are rated unsatisfactory, others are rated highly satisfactory. The sub-criterion for intended impacts assesses whether projects' intended impacts were achieved or are expected to be achieved, while the sub-criterion for contribution only considers those impacts for which the evaluation reports aimed to establish the project's plausible contribution. Differential impacts refer to the extent the impacts were inclusive or equitable and this assessment is more mixed with 40% of unsatisfactory projects among those which were assessed. Correlations among Impact sub-criteria is highly variable, e.g. the rating for intended and unintended results is uncorrelated.

Figure 13 | Impact criteria



Note: Percentages may not total 100% due to rounding.

Figure 14 | Sustainability criteria



Note: Percentages may not total 100% due to rounding.

Sustainability

The (expected) sustainability of projects is rated comparatively poor. Capacity and resilience development are mostly rated (highly) satisfactorily with relevant deficiencies identified in 25% of cases. Financial sustainability on the other hand is rated unsatisfactorily in about half the cases. Further, recommendations for follow-on funding – not always reflected in the assessments – may indicate that even more projects might lack financial sustainability. The extent to which the context is favourable for sustainability, is rated in most cases positively (69%). Correlation patterns between the sub-criteria further reinforce the fragmented nature of sustainability assessments: contextual factors are uncorrelated from other dimensions, and the relationship between financial and capacity-related sustainability is weak.

INFO BOX 3: INTERNATIONAL COMPARISON OF DAC PERFORMANCE

In this box, we present the most recent performance scores for KfW and AfDB, organizations assessing performance using a similar scale.[27]

The German Kreditanstalt für Wiederaufbau (KfW) reported that, for the period 2023-2024, 46% of projects were rated as successful and 38% as moderately successful, resulting in an overall success rate of 85%.[28] These ratings are based on independent ex-post evaluations conducted several years after project completion.

The African Development Bank (AfDB) reported that 89% of projects completed in 2022-2023 were rated satisfactory overall. By evaluation criterion, 99% were rated satisfactory for relevance, 68% for effectiveness, 75% for efficiency, and 88% for sustainability.[29] These ratings are based on self-assessments prepared by operational staff in project completion reports, independently reviewed and validated by AfDB's Independent Development Evaluation (IDEV).

Compared to these institutions, SECO-WE's results are broadly similar to AfDB's, though sustainability is lower. KfW's overall success rate - no DAC-level breakdown was available - is slightly below SECO-WE's aggregate performance. These comparisons should be read with utmost caution, as methodologies differ and margins of error can be large - for example, KfW's 85% has a confidence range of 77-92%.

[27] KEK-CDC (2019). *Sustainability Review*.
[28] KfW (2025). *18th Evaluation Report 2023–2024: Evaluate. Measure. Learn.*
[29] AfDB IDEV (2025). *Synthesis Report on the Validation of 2022-2023 Project Completion Reports and Expanded Supervision Reports*.

Internal versus external assessments

On average, there is little difference between internal and external assessments of projects across OECD-DAC criteria. In addition to mostly external evaluations, some projects are also assessed internally in project completion notes along the same dimensions of the OECD-DAC criteria. There are only minor differences in ratings between internal and external assessments for the criteria Relevance, Coherence, Impact and Efficiency. Only for Effectiveness and Sustainability the criteria differ and are on average rated slightly lower in the external assessments. For further details refer to Annex 8.

Availability of information for the assessment along the OECD-DAC criteria

As noted also under limitations (see Chapter 3), the availability and depth of assessment across evaluation criteria vary significantly across the reviewed reports, with some areas receiving more consistent attention than others. Relevance is generally well covered (92%), though not without gaps. Coherence (77%), particularly internal coherence within the Swiss international cooperation (IC) system (37%), is less comprehensively addressed. Effectiveness assessments are present in most evaluation reports (92%), but many sub-criteria are rather rarely addressed (25%). Efficiency is assessed in most evaluations (85%) but based on different sub-criteria (each 63-75%). Impact is the least assessed criteria (63%) and remains one of the most challenging criteria to assess due to limitations in the evidence base. Sustainability is covered unevenly across reports (81%), with notable information gaps (see also Figure 22 in Annex 9).

CASE STUDY INSIGHT 2: DETERMINANTS OF PERCEIVED EVALUATION QUALITY

Several key determinants of evaluation quality emerged from the case studies, taking the perspective of the users – and which are closely linked to the **context, process, and content of the evaluation**.

For many stakeholders, the quality of an evaluation depends heavily on the evaluator's understanding of the project's context and objectives. Those who demonstrate familiarity with both the local context and the project's goals are often seen as more credible. For example, in the case of the 'External Evaluation – Swiss Trade Policy and Export Promotion Project Vietnam' (WEHU 227), the evaluator's in-depth knowledge of the local context and their ability to communicate effectively with relevant stakeholders on both the Swiss and Vietnamese sides were identified as major strengths of the evaluation. Interviews, however, also emphasised the importance of evaluators having profound evaluation expertise, besides knowledge of the context and subject matter.

Evaluation quality, moreover, is often assumed through **established checks in the system**. Project managers, both from SECO and from SECO's partners, often referred to and underscored the importance of a robust internal system (within SECO or its partners) for reviewing and approving evaluations which ensures that methodological standards are met. Although they did not necessarily consider themselves capable of properly assessing the quality of the evaluation, they often referred to existing systems, trusting that these would highlight any weaknesses. Such systems included the assistance of a dedicated evaluation unit, the development and approval of clear terms of reference (ToR) for each evaluation, the selection of qualified contractors, and the involvement of internal project stakeholders or an evaluation subcommittee to review draft reports, means of data collection, interviewee lists etc.

> "I guess if the evaluation was splattered with mistakes, it would not have been taken up." (Partner)

However, some cases highlighted that, despite these checks, there can be situations where the evidence is uncritically accepted. To mitigate this, it is critical to ensure that evaluations undergo review and quality control, such as through tools like REQA or checklists (see also Chapter 6 on recommendations).

The perceived quality of an evaluation further was seen as closely linked to the **usefulness of its recommendations**. Stakeholders value evaluations that provide specific, actionable, and realistic recommendations (see also Insight 3). It was found that, for recommendations to be adopted, they must **resonate with stakeholders and align with their expectations**. Recommendations that affirmed stakeholders' perceptions or confirmed their intuitions were found to be more readily accepted. However, recommendations that were unexpected or 'surprising' could also be valuable if they were supported by robust evidence. As demonstrated by the 'Evaluation of the Colombia Más Competitiva Programme' (WEHU 226), while some of the evaluation's conclusions were not universally agreed upon, the discussion process helped to clarify their implications, and most recommendations were adopted for the next phase.

Finally, **participation in the evaluation process** is a critical factor in perceived quality. Evaluations that involve stakeholders at multiple stages – such as through regular updates, input on the ToR, and sharing of interim findings – were generally seen as valuable. Furthermore, the inclusion of multiple perspectives, such as from both internal and external experts, ensures that the evaluation's conclusions are grounded in diverse viewpoints. In all of the cases, the participatory process was emphasized, with stakeholders actively engaged throughout, ensuring that the evaluation remained relevant to their needs and that the findings were clearly understood.

> "An iterative and participatory process was used to improve the quality of the report and ensure the reliability of the data sources and the evaluator's understanding of the business model. A lot of work went into making the report usable." (Project manager)

Quality of Theories of Change and Logframes

Clear and well-articulated Theories of Change (ToC) and logframes are more than technical tools - they can be critical enablers of project effectiveness. Evidence from institutions like the World Bank shows that projects with high-quality M&E systems, including robust ToC and indicators, are significantly more likely to achieve stronger outcomes.[30] This highlights the importance of examining how such frameworks are used and reflected in evaluations.

While having a ToC is considered good practice, it is not mandatory at SECO-WE. An internal SECO-WE analysis of credit proposals revealed that in 2018, only 3% of projects included a ToC; by 2021, this had increased to 27%, signalling their growing uptake. Logframes, by contrast, are mandatory for all projects with a budget exceeding CHF 1 million.

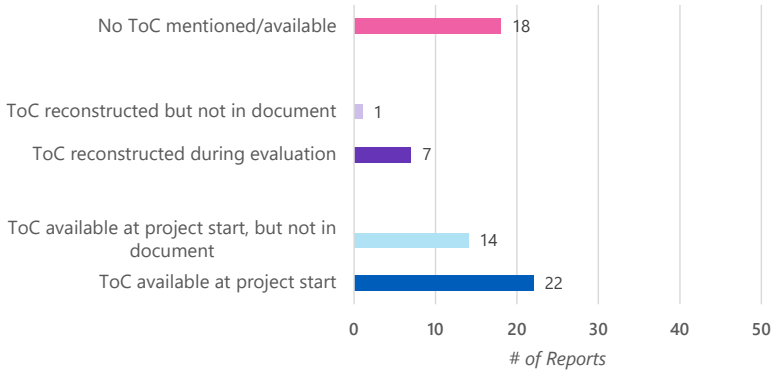
Against this backdrop, the following section explores how project-level ToCs and logframes are assessed in evaluation reports - with particular attention to their quality - and how they relate to both project performance and evaluation characteristics.

Project-level Theories of Change

Availability and Types

The content analysis shows that Theories of Change are inconsistently available and often insufficiently documented in evaluation reports. Across all 47 reports reviewed, information on the availability of a Theory of Change (ToC) was identifiable. In most cases (22), a ToC was reported to be available from the outset. However, in 14 of these 22 reports, evaluators referred to a pre-existing ToC but did not present it in the evaluation document itself, raising questions about accessibility and transparency. Notably, seven evaluations reconstructed a ToC as part of their process, and in one of these cases the reconstructed version was not included in the report. Nevertheless, in nearly 40% of cases (18 reports), no ToC was mentioned or documented at all. This distribution of cases highlights recurring shortcomings in documentation and transparency: either ToC are not included despite being referenced, or they are not available at all. From an evaluative perspective, the inconsistent inclusion of ToCs undermines the transparency of project logic and limits the possibility to assess how the ToC shaped the evaluation approach or findings.

Figure 15 | Availability (n=47)



The content analysis reveals a general lack of specificity of Theories of Change (ToCs) limiting their use to guide evaluation approaches. Out of the 29 reports that mention a project-level ToC, 15 did so without clarifying the structure or identifying a specific type - suggesting either a lack of formalisation or insufficient reporting. Among the remaining cases where a ToC was included in the document, graphical representations were most common (14 reports). Narrative ToCs and nested formats appeared far less frequently, each in only four reports. Interestingly, none of the reports featured a ToC with an "if–then–because" structure or an explicitly formulated hypothesis chain.

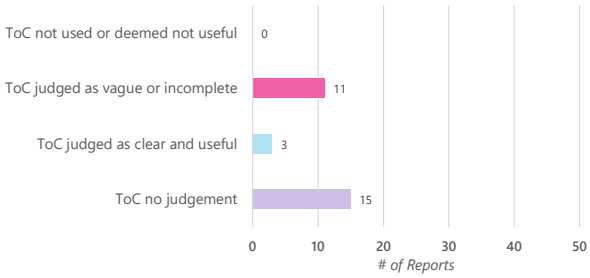
[30] World Bank Independent Evaluation Group. (2024). *The importance of monitoring and evaluation for World Bank project performance in eight graphs*; Raimondo, E. (2016). *What Difference Does Good Monitoring & Evaluation Make to World Bank Project Performance?* (7762; Policy Research Working Paper).

The ToC types were not mutually exclusive and occasionally appeared in combination. In cases where ToCs were presented graphically, their quality varied substantially. In some cases, graphical ToCs did differentiate between levels (e.g. outputs, outcomes), but often in a rudimentary fashion. This lack of detail limits the Theories of Change's potential to support meaningful causal reasoning or guide evaluative inquiry.

For evaluations mentioning to be following a theory-based approach, Theories of Change (ToCs) are often missing - but less frequently. While evaluations identified as theory-based always mentioned a Theory of Change, they still frequently failed to include it in the report itself. Of the 13 evaluations identified as theory-based, only 6 included the ToC in the document, while 4 referred to it without presenting it, and 3 had to reconstruct it during the evaluation. For other evaluations, less than half refer to a ToC in the document and most of those that do, do not include it in the document. Only 6 out of 35 evaluations not identified as theory-based include a Theory of Change in the report. This indicates that while theory-based evaluations do engage more often with ToCs, the persistent lack of documentation - even in these cases - points to a broader challenge of transparency and reporting quality.

Quality

Figure 16 | Theory of Change Quality (n=29)



Theory of Change (ToC) quality is rarely assessed in depth, and where it is, evaluations frequently identify conceptual and structural shortcomings. Across the 29 reports which reference a Theory of Change, half of the relevant cases (15) come without a judgement on the ToC quality - either because it was not assessed or because the evaluators themselves had to reconstruct it during the evaluation, which limited their ability to comment on its initial quality. Of the remaining 14 reports, only three explicitly judged the ToC as clear and useful. In contrast, eleven evaluations described it as vague, incomplete, or insufficiently grounded, often highlighting weaknesses in logic, assumptions, or articulation of contribution. Notably, none of the evaluations stated that a ToC was available but ultimately not used or found to be irrelevant.

The critical assessments of Theories of Change commonly point to a limited articulation of causal pathways. Several evaluations noted that assumptions between outputs, intermediate results, and long-term outcomes were either missing or not explicitly stated. For example, one report judged the ToC as incomplete because "the expected contribution of activities to outputs and outcomes is missing" ('External evaluation of the African Cities Lab in urban development project' (WEIN 94)), despite being presented alongside a logframe. Other evaluations criticised high-level or overly generic assumptions that lacked empirical grounding or specificity - especially in areas such as the link between project activities and systemic change (e.g. 'Impact Evaluation for Global Reporting Initiative' (WEHU 225), 'Independent terminal evaluation Global Quality and Standards Programme' (WEHU 216)). In some cases, key assumptions were simply found to be unrealistic or untested, undermining the credibility of the ToC as a framework for guiding implementation and learning.

Suggestions for Improvement

Even when problems were identified, only a few evaluations provided concrete suggestions for improvement. Among those, the most frequent recommendations included increasing the flexibility of the ToC to allow for iterative adaptation (4 reports), clarifying underlying assumptions (2 reports), and grounding the logic in more realistic contextual factors (3 reports). Remarkably, no report offered structured guidance on how to improve the Theory of Change's strategic or analytical use. This further reinforces the impression that while evaluators may note conceptual weaknesses, they often fall short of engaging the ToC as a tool for critically assessing and reflecting the project's underlying logic.

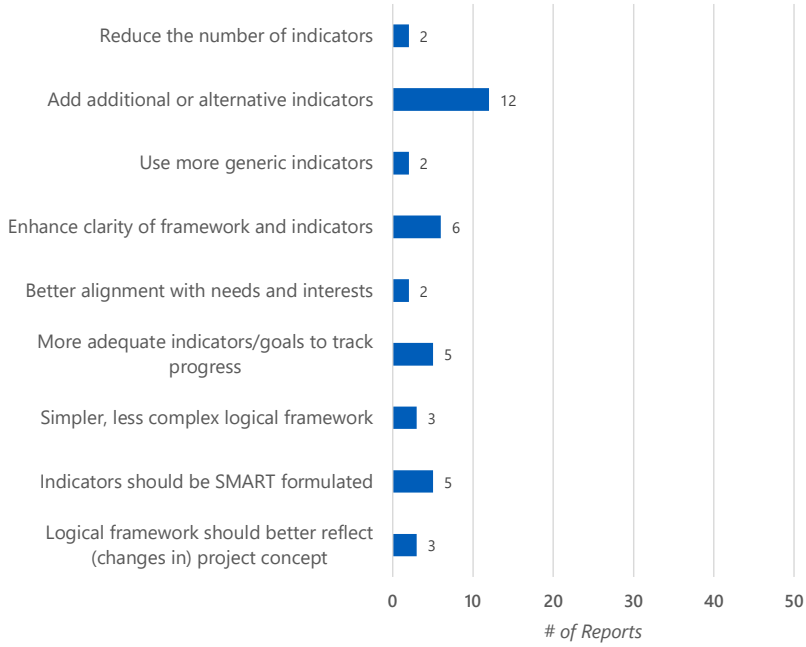
Logframes

Quality

Logframe quality is mixed, with most logframes assessed as inadequate due to unclear logic, weak assumptions, or unsuitable indicators. Across the 31 evaluation reports that assessed logframe quality, a majority (18) judged them to be deficient or inadequate. A key reason for this was that central assumptions, newly introduced outputs, or crucial contextual conditions were not sufficiently reflected in the frameworks. In other cases, logframes appeared too abstract or overly ambitious, offering little practical value for managing or assessing results. Common problems further included logframes that conflated different results levels (e.g. mixing outputs with outcomes), lacked coherence between activities and intended effects, or were simply too complex to navigate. However, almost half (13 reports) were judged as clear and adequate. For example, the 'Final Evaluation of For example, the 'Final Evaluation of the GovID programme' (WEMU 116) praised a clearly structured logframe with well-defined indicators, baselines, and verification mechanisms - updated regularly through internal and external data collection. Likewise, the 'External Mid-term Evaluation and outlook of Renewable Energy Skills Development (RESD) in Indonesia' (WEIN 92) found that the logframe was "relevant and encompass[ed] all necessary elements," making it a useful tool for implementation.

Indicator quality is a common weakness across evaluations and frequently undermines the quality of logframes. Indicators, as a central pillar of any results-based framework, were judged in 27 reports. Only 6 considered them relevant and measurable; in contrast, 21 found them deficient. Reported weaknesses included indicators that were not SMART (specific, measurable, achievable, relevant, time-bound), not linked to actual data sources, or not available at all.

Figure 17 | Suggestions to improve logframe (n=23)



Suggestions for Improvement

Evaluators frequently recommend targeted improvements to make logframes and indicators more useful, realistic, and adaptable to evolving project needs. Across 23 reports, evaluators offered concrete suggestions to improve the design and practical use of logframes and indicators - pointing to several recurring challenges. The most common recommendation, found in 12 evaluations, was to introduce additional or alternative indicators to better reflect project outcomes. Often, this was linked to concerns about incomplete or overly narrow measurement systems that failed to capture meaningful change. Six reports advocated for clearer articulation of indicators and overall framework logic, while five highlighted the need for indicators that are better aligned with project progress or formulated in line with the SMART criteria. These suggestions reflect concerns that current frameworks may not provide a reliable basis for monitoring or learning. Beyond indicator content, structural issues were also raised. In three evaluations, the logframe was criticised for failing to reflect updates or changes in the project design, calling for revisions to

ensure internal consistency. Another three reports noted that the logframes were overly complex and recommended simplification to improve usability in practice. These issues echo the deficiencies in logframe quality regarding limited clarity and adaptability mentioned above. Several evaluations also addressed the contextual fit of the frameworks: two reports recommended stronger alignment with local needs and stakeholder priorities, while two others proposed using more generic indicators to allow comparability across contexts. In a few cases, evaluators suggested reducing the overall number of indicators to focus monitoring efforts and prevent overload.

Quality and Project Performance

Explicit links between Theory of Change quality and project effectiveness are rarely explored in the evaluation reports, and overall inconclusive. Across the 47 reports, only three evaluations discussed a potential relationship between the quality of the ToC and project effectiveness. One evaluation suggests that a strong, coherent Theory of Change may contribute to effective implementation and meaningful results, especially when assumptions are realistic and causal pathways well-articulated, as this clarity helps to align interventions with intended outcomes and supports strategic decision-making during implementation. One evaluation – the 'Midterm Evaluation of the ITC MENATEX Programme' WEHU 213) – explicitly mentions that an inadequate Theory of Change may have hindered project performance and recommends strengthening it to improve operational efficiency. One other – 'Independent terminal evaluation Global Quality and Standards Programme' (WEHU 216) – describe the relationship as unclear or mixed, for instance due to vague logic or weak alignment with expected results. Importantly, none of the evaluations offered in-depth analysis of the specific design features that made a ToC more or less effective in practice. This lack of systematic reflection stands in contrast to the central role that ToCs are assumed to play in evaluation. Where the link is addressed, observations remain largely anecdotal and underexplored.

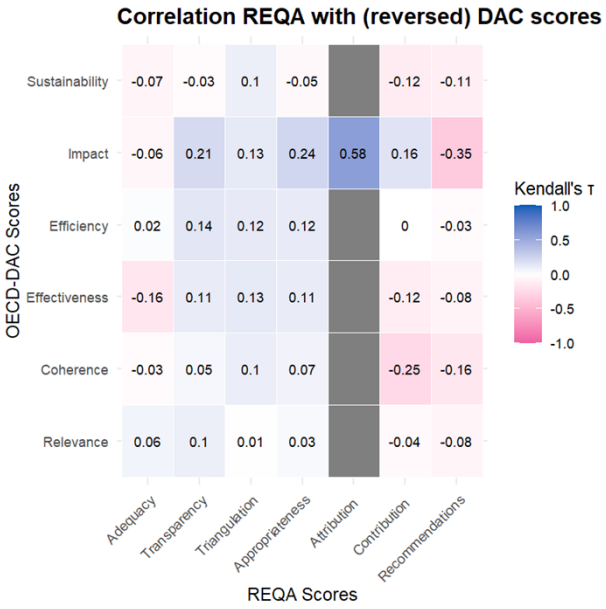
Similarly, few evaluations explicitly explore the link between logframe quality and project effectiveness indicating that vague logframes and a focus on activities hinder effectiveness. Across the 47 reports, 6 included reflections on whether and how the quality of the logframe influenced project outcomes. Of these, 2 reports described a positive connection, suggesting that a well-structured logframe supported implementation and allowed for more strategic decision-making. For example, the evaluation of a regional programme noted that both project- and programme-level reporting were "well detailed" and that the logframe enabled "satisfactory tracking of activities at outcome and output levels," thereby facilitating effective steering ('GPIPR External Mid-Term Evaluation' (WEHU 211)). Similarly, 4 evaluations suggest that vague or overly activity-focused logframes hindered effectiveness. In the 'Mid-Term External Evaluation of ECOFEL Program' (WEMU 113), evaluators argued that clearer outcomes and indicators could have improved strategic guidance and encouraged more systematic reporting. Similarly, the 'Midterm Evaluation of the ITC MENATEX Programme' (WEHU 213) criticised a monitoring approach that focused heavily on activity delivery, while lacking tools to assess quality or stakeholder feedback thereby making it difficult to correct course. When looking at the overall DAC score for effectiveness in these six evaluations, the analysis shows no pattern. A detailed frequency table can be found in Annex 11.

The intersection of DAC scores with coded assessments of Theory of Change quality reveals possible associations between stronger Theories of Change and better project performance. To explore this, the DAC (sub-)criteria *Relevance – Sub-criterion: Quality of Design*, *Effectiveness*, *Effectiveness – Subcriterion: Attainment of Objectives*, *Impact*, and *Impact – Subcriterion: Intended Impacts* were cross-analysed with the qualitative assessments of ToC quality identified in the content analysis. The results suggest that evaluations featuring a clear and useful ToC more often score higher in the DAC criteria for relevance and effectiveness. In contrast, evaluations with vague or incomplete ToC tended to be associated with lower scores. For the impact ratings, no such patterns are apparent from the data, similarly if the ToC quality was not assessed. Notably, the highest DAC ratings were almost always for projects for which the evaluation report reconstructed the Theory of Change, pointing to a link between quality of the evaluation report and OECD-DAC ratings.

Cross-analysing logframe quality with OECD-DAC scores does not reveal a relationship of logframe quality with project performance but it predicts whether impact was assessed. To explore the relationship between logframe quality and project performance, the DAC (sub-)criteria *Effectiveness, Effectiveness – Subcriterion: Attainment of Objectives, Impact*, and *Impact – Subcriterion: Intended Impacts* were cross-analysed with qualitative assessments of logframe quality. The data suggests no relationship with ratings of Effectiveness or Impact. However, while in most cases when the logframe was judged to be clear and adequate, the evaluation assessed the project's impact. This was the case for only one-third of the evaluations with a deficient or inadequate logframe. This mirrors the finding from the REQA that impact assessments frequently are based primarily on an analysis of project indicators and often indicate the lack of reliable data or impact logic as a limitation to their assessment.

A complementary analysis of REQA and OECD-DAC scores further suggests that higher-quality evaluations are associated with better project performance. Figure 18 displays the correlation between REQA and OECD-DAC scores.[31] Projects with higher REQA scores for analysis related criteria tend to be rated more favourably on impact, efficiency, and effectiveness, but the correlation is weak. In the case of impact, the correlation is primarily driven by the sub-criteria for contribution to intended impacts, unintended impacts, and differential results. This association is partly structural: the assessment of contribution explicitly considers the quality of evidence - a central concern in REQA. Similarly, the identification of unintended and differential outcomes typically requires stronger data and more rigorous analysis, both of which are characteristics of higher-quality evaluations. Note that these correlations do not imply causations. In particular, it is unplausible that higher-quality evaluations cause higher performance of the evaluated project, as measured by those same evaluations. However, these correlations indicate that high impact ratings are not necessarily associated with lower-than-average evidence quality.[32] That said, this does not imply improving evidence quality of evaluations would lead to higher performance ratings as selection effects may play a significant role. For instance, greater effort or resources might be invested in evaluations intended to demonstrate success rather than verify failure.

Figure 18 | Correlation between REQA and DAC scores



[31] The figure presents the correlations between REQA scores and reverse-coded OECD-DAC criteria. A positive correlation indicates alignment in the direction of judgment.
[32] But recall that the overall evidence quality across evaluations, especially on the impact level and with regards to causal claims, was found to be generally low based on both the REQA and the content synthesis.

CASE STUDY INSIGHT 3: TIMELINESS AND PRACTICALITY – KEY FACTORS FOR THE USEFULNESS OF EVALUATIONS

The practical usefulness of evaluations depends on two main factors, as identified in the case study data: the timeliness of the evaluation, and the quality of its recommendations. Firstly, the **timing of an evaluation** is crucial as it provides input for the design of subsequent project phases. For example, in the case of the 'Evaluation of the Colombia Más Competitiva Programme' (WEMU 226), the evaluation was particularly impactful as it occurred at a critical juncture: the project was transitioning from its second phase to its final phase while the Swiss Development Cooperation (SDC) was scaling down its involvement in the country. This timing enabled the evaluation to inform critical decisions concerning the project's shift in focus and the restructuring of responsibilities, ensuring the transition remained aligned with the project's objectives and broader diplomatic priorities. The evaluation's findings, particularly regarding the project's long-term sustainability and focus on strengthening economic ties between Switzerland and Colombia, were politically opportune and valuable in helping to refine the project's direction and make its outcomes more relevant to stakeholders and the evolving political context. A similar conclusion was reached based on four interviews conducted to assess the usefulness of the independent climate evaluation. At the time of the evaluation, SECO-WE was rethinking its climate engagement and commissioned the evaluation to inform the development of a new climate policy and subsequent guidelines. The evaluators and internal stakeholders emphasised the importance of momentum and institutional openness. The evaluation thus benefited from a 'policy window', allowing evidence to be adopted.

> "There was an enormous interest and push to do more on climate… Politically, there was momentum. And internally, SECO really wanted to know: 'Are we doing this right?'" (Evaluator)

Similarly, the cases demonstrated that IMF evaluations (e.g. mid-term evaluation of the programme "Revenue Administration and Public Financial Management Reform in Southeast Europe" (WEMU121) are intentionally planned as mid-term evaluations to inform the design of the next phase. It was also noted that these evaluations are valuable for fundraising efforts, as they demonstrate to potential funders the relevance and potential effectiveness of investing in the next phase.

> "It was a critical time, primarily because it coincided with our preparations for the next phase of the project. This allowed us to ensure that all the recommendations from the evaluation were incorporated into the design of the new phase, improving efficiency and effectiveness." (Partner)

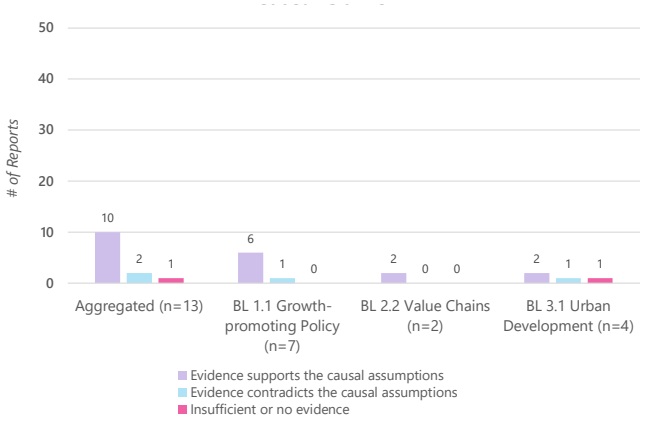
Secondly, the perceived usefulness of evaluations is significantly influenced by the **practicality of their recommendations**. Effective evaluations clearly communicate evidence, offering concrete, actionable recommendations tailored to the project context. For instance, in the case of the evaluation of the Extractive Industries Transparency Initiative (EITI), Phase IV (WEMU 112), the recommendations were recognised as being highly specific and well-supported by evidence, prompting in-depth discussion within the project board and resulting in their adoption. Similarly, the "External Evaluation – Swiss Trade Policy and Export Promotion Project Vietnam" (WEHU 227) was said to have had a positive influence on the ongoing phase and to have provided direction for the next phase, highlighting the differences in implementation approaches between the International Trade Centre and the Ministry of Industry and Trade. In some cases, interviews revealed that recommendations are subject to critical scrutiny, as demonstrated in the case of the "Final Evaluation Report – EBRD-SECO ITCP Early-Stage Assessment" (WEIF 136), where a recommendation regarding the potential effects of brain drain lacked sufficient evidence to be fully convincing. While methodological considerations, such as those employed by the REQA, are one factor in determining the usefulness of evaluations, it is ultimately the **actionable, context-specific insights that provide clear guidance for project refinement and improvement** (see also Insight 2).

Evidence base of Impact Hypotheses

Evidence-based programming has gained traction at SECO-WE. In this section we aim to lay out the causal evidence from evaluation reports against three selected Business Line Impact Hypotheses, namely, 1.1 Growth-promoting policy, 2.2 Integration into Value Chains, and 3.1 Urban Development.

Causal Claims

Figure 19 | Causal Claims



Causal assumptions are rarely made explicit or explored. Only one third of reports (13 out of 30) reflected how project activities were expected to contribute to outcomes under the respective Business Lines. Among the reports that included causal claims, seven were linked to Business Line 1.1 Growth-promoting policy. Business Line 2.2 Integration into Value Chains and Business Line 3.1 Urban Development were less frequently represented in the reports that included causal claims - each with only one or two cases. Even among these 13 reports only 9 included a causal analysis as identified in the REQA. This relatively small number of reports engaging with causality in a systematic way suggests a limited examination of the underlying logic of interventions.

Where causal claims are tested, they are more often confirmed than contradicted across Business Lines, with Growth-promoting policy projects most frequently engaging with causal logic. Among the reports that included causal claims, seven were linked to Business Line 1.1 Growth-promoting policy. Six of these confirmed the causal assumptions, while one identified a contradiction. Business Line 2.2 Integration into Value Chains and Business Line 3.1 Urban Development were less frequently represented in the reports that included causal claims - each with only one or two cases - but showed a similar pattern: causal assumptions were generally supported or yielded mixed results. While this suggests that when causal pathways are articulated, they are mostly confirmed, it needs to be noted that the extent to which they are substantiated by evidence and analysis is often limited.

Contextual Factors

The evaluations frequently report that Covid-19, political factors, and the availability of human resources and institutional capacities influenced project success. The Covid-19 pandemic was frequently mentioned in 10 out of 30 reports. It caused delays, disrupted stakeholder engagement, and hindered in-person activities such as training or networking - affecting both timelines and outputs. Equally commonly mentioned were political factors (10 reports), which included shifting priorities, institutional instability, and weakened alignment with national strategies, often complicating implementation or reducing stakeholder commitment. A third relevant factor was the availability of human resources and institutional capacities (6 reports), such as high staff turnover or insufficient technical capabilities, which limited implementation continuity and long-term anchoring of results.

Business Line Thematic Analysis

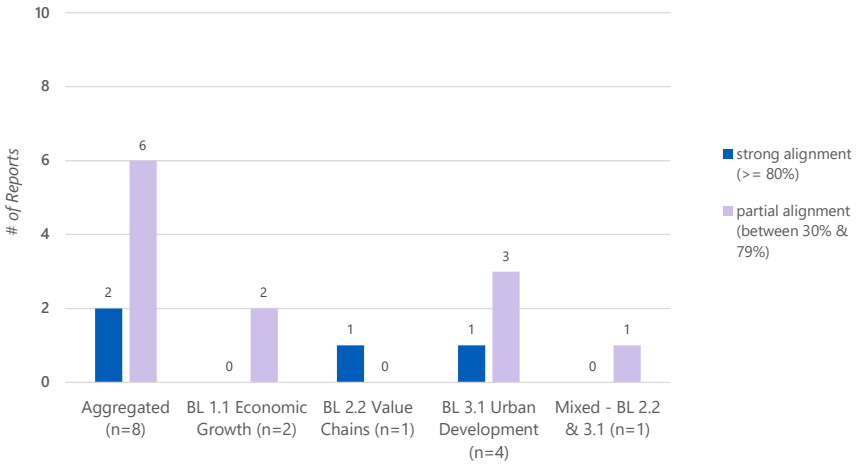
Across all three Business Lines, the thematic analysis shows a consistent reporting focus on activities and outputs, while outcome- and impact-level results are less systematically addressed. Evaluations tend to describe what was done rather than what was achieved, with the depth of analysis diminishing along the results chain. In Business Line

1.1 (Growth-promoting Policy), institutional measures dominate, but broader outcome diversity and impact linkages remain limited. Business Line 2.2 (Integration into Value Chains) shows the weakest coverage beyond outputs, with notable gaps in key themes like producer resilience. Business Line 3.1 (Urban Development) reflects more outcome and impact-level content, particularly regarding service access, but underrepresents themes related to resilience and environmental sustainability. Overall, the evidence suggests a general pattern of operational reporting with limited tracing of long-term or systemic change. Detailed results can be found in Annex 10,

Alignment of Project-level Theories of Change with Business Line-level Theories of Change

Due to the limited availability of project-level Theories of Change, the assessment of their alignment with the corresponding Business Line Theories of Change could only be conducted for a small number of evaluation reports. Out of all reviewed reports, only eight offered enough detail to examine this relationship systematically. Alignment refers to the extent to which the project-level Theory of Change mirrors the logic, causal pathways, and strategic objectives articulated at Business Line level.

Figure 20 | Alignment of project Theories of Change with Business Lines



In six of eight cases with sufficient details on the project's Theory of Change, the alignment was partial. This means that the project Theory of Change reflected some elements of the Business Line framework (between 30% and 79%). For instance, in Business Line 1.1 Growth-promoting policy, the 'Independent Evaluation of the Extractive Industries Transparency Initiative' (WEMU 112) included several relevant components such as institutional reform and improved governance, but did not fully address broader macroeconomic or fiscal outcomes. Similarly, a project bridging Business Line 2.2 Integration into Value Chains and Business Line 3.1 Urban Development the 'Independent Evaluation of Mobility and Logistics (MOLO)' (WEIN 86) showed selective correspondence, aligning with certain value chain outcomes and urban development priorities without capturing the full scope of either Business Line.

Strong alignment, defined as at least 80% overlap, was found in two of the eight reports. These reports demonstrated a high degree of coherence between project- and Business Line-level logics across all levels of results, as assessed through the content analysis. In Business Line 2.2, the 'Evaluation of the Sustainable Trade Initiative's work on living wages' (WEHU 226) clearly aligned with the Business Line focus on improved working conditions, producer integration, and competitiveness. For Business Line 3.1, the 'End-of-program external Evaluation of the activities carried out in Colombia and Peru under the Umbrella of the IFC LAC Sustainable Cities Program' (WEIN 90) explicitly mapped urban service delivery, innovation, and sustainability back to the impact logic of the Business Line Theory of Change.

CASE STUDY INSIGHT 4: CHALLENGES IN ESTABLISHING CLEAR CAUSAL LINKAGES – THE ROLE OF THEORY OF CHANGE AND DATA QUALITY IN PROJECT EVALUATIONS

One of the major challenges identified in the analysis relates to the difficulty of generating convincing evidence of clear causal linkages through evaluation. The case studies provide further insight into this issue, highlighting the **absence of a well-defined Theory of Change (ToC) or intervention logic, as well as inadequate monitoring data**. Several evaluations noted the poor quality of logframes and ToCs, and recommendations were often made to improve them – or the evaluations themselves contributed to the development of a better ToC. For instance, evaluators have observed that **projects often lack fundamental elements**, such as clearly defined outcomes and a properly structured logframe used for project management. Without clear indicators, the evaluation process becomes problematic. The absence of foundational elements, such as robust logframes or ToCs, makes it difficult to evaluate outcomes effectively and determine causal relationships between inputs and results.

> "All that sophisticated stuff [relating to contribution or attribution analysis] is just built on sand unless you get proper logframes, results frameworks." (Evaluator)

In other cases, data quality was more pronounced. For instance, the IMF's 'Mid-term Evaluation – Revenue Administration and Public Financial Management Reform in Southeast Europe' (WEMU 121) faced difficulties with data availability, and sample sizes were found to have been too small or not homogenous enough to conduct in-depth quantitative or contribution analyses. In many cases, the evaluators also struggled with incomplete or deficient monitoring data, illustrated by the following statement. In these cases, evaluations often had to be more flexible, as the available data and frameworks were insufficient for rigorous causal analysis.

> "The data was even worse than normal... the logframe was pretty weak, there was no ToC, many technical, IT-related problems... the indicators are only at the outcome level; and often they were not really relevant". (Evaluator)

However, it was also highlighted in many cases that **evaluation offered direct contributions to improving the project's foundations**, like a clear(er) ToC. For example, in the case of the 'External Evaluation – Swiss Trade Policy and Export Promotion Project Vietnam' (WEHU 227), one of the stated goals of the evaluation was to review the project's ToC, which had significant flaws. The evaluation process involved collaborative work with the project team and partners to clarify the goals and strengthen the ToC. As one evaluator stated:

> "The intention was clear: to look at the project's theory of change, especially regarding future goals... the first step was to work with the project and partners to figure out what the goal actually was." (Evaluator)

In conclusion, the difficulty of establishing clear causal linkages is often due to a lack of solid foundations in terms of objectives, indicators and data collection. While evaluations have highlighted these gaps, they have also provided practical recommendations for improving the ToCs and logframes. This has enhanced the clarity of causal linkages (e.g. for a follow-up phase) and may enable a more accurate assessment of project impact and effectiveness in the future.

5 Conclusions

This report has examined the evidence quality and synthesized the contents of project evaluations of SECO-WE supported projects. It has also assessed project performance based on these evaluations. The following section summarizes the key findings, offering an overview of the insights gained and highlighting the identified strengths and challenges.

Conclusions from the Rapid Evidence Quality Assessment (REQA) of evaluations

The analysis of REQA scores highlights **considerable room for improvement in the overall quality of evaluations, particularly across key dimensions such as Transparency, Triangulation, Appropriateness, and Causal Analysis**. With the exception of Adequacy – where 92% of evaluations were judged appropriate for their stated objectives - most criteria scored at or just above a basic level, with few evaluations achieving very high standards. Similar to these findings for SECO-WE's evaluations, other assessments and reviews of evaluation practices also highlight methodological weaknesses in evaluations carried out by a range of agencies and donors.

Transparency remains a critical weakness, with reports frequently omitting key methodological details and failing to reflect on data limitations. Similarly, while triangulation methods are often used, perspectives are rarely contrasted or reconciled. Appropriateness shows somewhat stronger performance, yet recurring gaps in data quality and analytical rigor and the lack of causal analysis compromise the credibility of evaluation report findings.

The interrelationship between criteria may underscore the importance of a holistic approach to quality enhancement. Strong performance in one area – such as Transparency – is generally associated with higher scores in others, such as Triangulation and Appropriateness. This reinforces the need for integrated quality assurance measures.

Overall, while the findings indicate incremental progress between 2023 and 2024, sustained efforts are needed to advance evaluation quality from basic to good or exemplary levels. The limited quality of evaluations is a global concern, and a systematic review such as REQA can provide a foundation for improvement.

Conclusions from the OECD-DAC criteria assessment across evaluated projects

The analysis of OECD-DAC criteria across the evaluated projects reveals an **overall satisfactory level of performance, albeit with notable variation across criteria**. This must be cautiously interpreted since the evaluations used to score performance are largely of basic quality (REQA). Relevance, Coherence, and Impact are the strongest criteria with more than 85% of evaluated projects being rated (highly) satisfactory. Effectiveness and Efficiency are also strong but with more unsatisfactory and fewer highly satisfactory ratings, while Sustainability continues to represent the weakest area, with 41% of evaluated projects being rated as unsatisfactory.

Relevance is generally well substantiated, with strong alignment to needs and context, though design weaknesses (e.g., lack of impact logic) persist. Coherence is positive overall, but gaps remain in the analysis of internal compatibility, particularly in evaluations not commissioned by SECO. Effectiveness is weakened by inconsistent assessment of unintended or differential effects, despite satisfactory ratings in achievement of objectives. Efficiency ratings are mostly satisfactory but sometimes constrained by weak data and superficial analyses. Impact assessments remain methodologically limited and inconsistently applied, with little exploration of unintended or differential impacts. Sustainability is the most unevenly assessed and compared to the other criteria most poorly rated criterion, particularly in terms of financial sustainability.

The relatively minor differences observed between internal project completion notes and external evaluations, particularly the **absence of strong systematic upward bias in internal assessments**, may indicate a degree of robustness in internal assessment systems. While external evaluations tend to be slightly more critical, especially regarding Effectiveness, the overall consistency suggests that internal assessments may provide a reasonably reliable account of project performance. This raises questions about the extent to which performance against DAC criteria from

external evaluations adds value for accountability purposes, though further analysis would be required to substantiate such a conclusion.

Finally, the availability and quality of information for each criterion varies widely across reports. While most evaluations allow for at least a basic assessment of core DAC criteria, dimensions such as unintended effects, contribution to impact, or financial sustainability are frequently underexplored or supported by weak evidence. Overall, while project performance remains largely satisfactory, there may be a need to strengthen evaluation practices – particularly around Impact, and methodological depth.

Conclusions from the analysis on the quality of Theories of Change and logframes

The analysis shows that Theories of Change (ToCs) are referenced in a relatively small share of evaluation reports. Furthermore, their quality, structure, and documentation remain inconsistent. Many ToCs are only vaguely formulated, lack a clear articulation of assumptions or contribution logic, or are entirely absent from the evaluation report. Even when included, few are assessed critically in terms of their clarity or usefulness. These identified deficiencies limit the ToCs potential to support causal reasoning, evaluation design, and eventually guide (future) project design and implementation.

Similarly, while logframes are usually available, they often fall short in quality. A significant number of evaluations highlight unclear logic, weak differentiation between results levels, and insufficient or inappropriate indicators. In particular, many indicators lack SMART quality or are disconnected from meaningful data sources, reducing their value for monitoring progress or capturing change. While some evaluations acknowledge well-designed logframes that support strategic steering, these remain the exception.

Across both instruments, few evaluations explicitly link quality to project effectiveness and such connections, when made, remain anecdotal. However, projects for which the Theory of Change is judged to be well-articulated tend to receive higher scores for relevance and effectiveness, and higher quality logframes appear to support more meaningful impact assessments. Taken together with a weak correlation of REQA and DAC scores, this indicates a (weak) relationship of Monitoring & Evaluation quality and project performance, or at least the possibility to demonstrate performance.

Overall, these findings point to opportunities. On the one hand, strengthening the design and use of results logics could contribute to more effective implementation, and thus higher performance. On the other hand, a more systematic treatment of ToCs in evaluations could support clearer accountability and intensify learning for results, particularly at the impact level.

Conclusions from the analysis of the evidence base of SECO's Business Line Impact Hypotheses

Corroborating similar findings in the REQA section, the evaluation evidence reveals a consistent gap in the articulation and testing of causal assumptions which underpin Business Line Impact Hypotheses. Only a minority of reports explicitly addressed how project activities were expected to lead to outcomes, and even fewer systematically analysed these assumptions. While causal pathways, when examined, were generally confirmed, this finding is weakened by the limited number of evaluations engaging with causality in a structured way. The lack of robust causal analysis restricts the ability to assess the effectiveness of interventions and limits learning on what works and why.

Contextual factors like political shifts, institutional capacity, and Covid-19 disruptions frequently shaped implementation across all Business Lines, underscoring the need to take into account external influences.

Across all Business Lines, evaluations tend to focus on activities and immediate outputs, with significantly less emphasis on outcomes and impacts. This operational focus limits visibility into longer-term changes and weakens the

ability to assess strategic contributions. While Business Line 1.1 shows relatively stronger coverage across results levels, Business Lines 2.2 and 3.1 often lack consistent reporting on key thematic priorities at the outcome and impact level. This suggests a need to strengthen the results orientation of evaluations beyond delivery metrics.

Alignment between project-level and Business Line-level ToCs was only systematically assessed in a small number of evaluations. Where examined, most cases showed (partial) alignment, with project logic capturing selected elements of the respective Business Line framework. The limited availability of detailed project-level ToC however constrained this analysis, underscoring the importance of ensuring that projects articulate their logic in a way that supports both strategic alignment and evaluative insight.

6 Recommendations

This Performance & Quality Report has highlighted that while the overall performance of the SECO-WE's evaluated projects is satisfactory, challenges remain in areas such as sustainability, impact assessment, and the quality of projects' results frameworks. This report has also identified several areas in which the quality of SECO-WE's project evaluations could be improved, particularly with regard to transparency and causal analysis. Evaluations often fail to consistently assess the pathways of interventions' causes, which limits their ability to inform learning and decision-making. Strengthening the methodological rigour of evaluations along with improving the quality of project's results frameworks could enhance the evaluations' usefulness for a multitude of purposes, including programme planning.

Recommendation 1: Strengthen commissioning and quality assurance, specially of causal evaluations

It is recommended that WEQA actively promotes causal evaluations, and that WEOP increases their share by commissioning more evaluations with causal analysis to strengthen the evidence base for programme effectiveness and impact. These evaluations should, at a minimum, apply a recognised theory-based approach which appropriately tests causal assumptions and contribution claims. When commissioning causal evaluations, WEQA and its backstopping mandate should be closely involved in the conception and design phases to ensure methodological rigour and, if relevant, alignment with SECO's overarching impact hypotheses. Adequate financial resources should be allocated to such evaluations to enable robust designs and sufficient data collection.

Rationale: The REQA findings and analysis of the quality of ToCs underscore significant gaps in the application of causal analysis across evaluations, with only a minority systematically engaging with causal assumptions or testing alternative explanations. This methodological shortfall directly limits the credibility of conclusions about programme effectiveness and impact since such conclusions inherently involve causal reasoning. Similarly, weak or absent ToCs and inconsistently structured logframes restrict evaluators' ability to analyse causal pathways and assess contribution claims with rigour. Strengthening the commissioning of causal evaluations, especially those that apply recognised theory-based approaches, could address these deficits by ensuring systematic testing of results logic, consideration of contextual factors, and corroboration of claims with robust evidence. Such a shift could enable more credible attribution or contribution analysis, improve learning on what works and why – also beyond the individual project if linked for instance to SECO's impact hypotheses – and close a major gap identified in the REQA analysis

Addressed to: WEQA and WEOP

Priority: High

Recommendation 2: Enhance transparency and completeness of evaluation reporting

It is recommended that WEQA updates its evaluation standards to require higher levels of transparency in reporting, including clearer guidance on report structure, length, and the level of methodological detail expected, including an assessment of data quality and the inherent limitations of the evaluation. Commissioners (WEOP programme managers) should explicitly require full transparency in the Terms of Reference, ensuring that evaluators make available all relevant documentation necessary to fully understand the evaluation findings and limitations. Evaluators

should comply by providing and referencing supporting evidence, the used or reconstructed results logic (Theory of Change and/or logframe), methodological annexes, a critical appraisal of data quality and limitations, and any additional documentation required to interpret the results and conclusions.

Rationale: Transparency emerged as one of the weakest REQA dimensions, with frequent omissions of methodological details, incomplete reflection on data limitations, and inadequate referencing of supporting evidence. This undermines both credibility and usability of findings, particularly when conclusions are not fully traceable to data or analytical processes. The lack of accessible documentation, including reconstructed results logics, methodological annexes, and critical appraisals of data quality, hampers replication and peer review. Case studies further indicate that the level of trust placed in evaluation findings is sometimes disconnected from the underlying quality of evidence; without early and explicit quality controls, there is a risk that decision makers may base choices on results that are methodologically weak, insufficiently substantiated, or potentially misleading. Enhancing transparency standards and requiring full disclosure in evaluation reporting directly responds to this weakness and could foster greater trust in findings and enable triangulation.

Addressed to: WEQA, WEOP, and evaluators

Priority: High

Recommendation 3: Strengthen M&E capacity and resourcing

It is recommended that WEQA and WEOP invest further in building the Monitoring & Evaluation capacity of WEOP programme managers, with particular emphasis on strengthening their ability to design and use robust Theories of Change and logframes. WEOP should ensure that projects are adequately resourced for high-quality M&E from the outset, if possible, within dedicated budgets/plans for data collection, learning, and adaptive management. To support the M&E capacity of WEOP programme managers, it is further recommended to review and update WEQA's evaluation report checklist for programme managers against the REQA criteria and findings of this report, and to emphasise the importance of substantial evaluation experience -– alongside contextual and thematic expertise -– when selecting evaluators, to ensure they are both familiar with and able to comply with methodological standards for evaluations.

Rationale: The analysis of ToCs and logframes points to weaknesses in the articulation, structure, and use of results frameworks, with limited integration into evaluation design and minimal influence on learning or accountability. Poorly formulated ToCs, unclear assumptions, and inadequate indicators reduce both implementation effectiveness and evaluative insight. This is compounded by the finding that gaps in analytical rigour sometimes originate in weak results logic at the project level. Strengthening M&E capacity, particularly in designing and applying robust ToCs and logframes and in commissioning high-quality evaluations (see also recommendation 1), could address these structural weaknesses, support more strategic steering, and create the conditions for higher-quality evaluations capable of more credibly linking interventions to outcomes and impacts.

Addressed to: WEQA and WEOP

Priority: Medium

Recommendation 4: Explore unintended effects through portfolio-level evaluation

It is recommended that WEQA (or a WEOP section) consider commissioning a portfolio evaluation focused on identifying and analysing unintended effects of SECO-WE interventions - both positive and negative. This could be achieved by applying approaches suited to uncovering emergent and unplanned outcomes, such as Outcome Harvesting, Most Significant Change or similar complexity-aware evaluation methods.

Rationale: The OECD-DAC criteria analysis highlights inconsistent and often non- or superficial treatment of unintended or differential effects, with evaluations tending to focus narrowly on activities and outputs. Impact assessments are methodologically limited, and especially unintended or emergent outcomes are underexplored despite their potential to inform adaptive management and strategic positioning. Complexity-aware methods like Outcome Harvesting or Most Significant Change could systematically surface these dimensions, particularly at a thematic or portfolio level, where patterns and cross-cutting insights may emerge. Addressing this gap explicitly through eval-

uations designed for this purpose could therefore strengthen learning and enhance the strategic value of evaluations.

Addressed to: WEQA

Priority: Low

Recommendation 5: Assess the comparative value of external and internal DAC ratings

It is recommended that WEQA assesses the comparative added value of DAC ratings based on external evaluations relative to internal project completion notes, with particular attention to their respective contributions to accountability, learning, and methodological rigour. Given the relatively minor differences between internal and external ratings, and the absence of systematic upward bias in internal assessments, such an analysis would clarify whether resources devoted to external DAC scoring yield proportionate benefits for accountability, or whether their primary value lies in other dimensions such as learning and causal evidence, in line with a potentially more focussed approach (as opposed to covering all DAC criteria). The results of this assessment – implemented with the next biennial performance and quality report – should inform strategic decisions on the optimal mix of internal and external evaluation modalities to maximise utility and cost-effectiveness.

Rationale: Findings from the DAC criteria analysis suggest relatively minor differences between internal completion notes and external evaluations, with no strong systematic upward bias in internal ratings. At the same time, most evaluations were found to focus on a selection of DAC criteria or specific aspects, meaning that a complete assessment of all DAC criteria based on a single evaluation is usually not possible. This raises a legitimate question about the marginal accountability value added by external DAC scoring. A structured examination of the comparative value of these two assessment approaches could clarify whether external evaluations' main contributions lie elsewhere -– such as in methodological rigour, independent perspective, informing steering or learning.

Addressed to: WEQA

Priority: Low

ANNEX

Biennial Performance Report 2023-24

1 Evaluation Matrix

Overall objective of the BPR: Bring together and synthesize different information from SECO-WE's M&E system to showcase what was achieved, what worked well, what less so, and why.

Table 1: Evaluation matrix

Evaluation question	Assessment criteria	Data sources, data collection, data analysis	Limitations
1. What is the quality of evidence provided in external project and independent evaluations? Which areas require improvement?	REQA, additional coding on methodologies and approaches	Source/Collection: • 48 external project evaluation reports • Interviews (primary data) for triangulation Analysis: Descriptive statistics; "spider diagrams"	• Focuses on the final report only, not the ToR or inception report. • What is not written cannot be assessed. • Emphasizes methods; criteria such as usefulness and participation are only considered to a limited extent
2. How do externally and internally evaluated projects perform against the OECD DAC evaluation criteria, and what explains variation in performance?	OECD DAC rating methodology	Source/Collection: • 23 external evaluation reports 2024: to be rated by Syspons • 25 external evaluation reports 2023: rated by KEK • 60 completion notes 2023 and 2024: rated by programme managers • Interviews (primary data) for triangulation • Standard Indicator Data Analysis: • Descriptive statistics: data aggregated upon to the 6 DAC criteria and not summarised into one rate • Content analysis of reasons for ratings (justification column)	• Not all projects apply all DAC criteria • The decision for a DAC criterion is based on knowledge gaps (learning) and steering/accountability considerations. • Comparison between internal and external evaluations only partially possible, because they mostly cover different projects (only 12 projects overlap) • No generalisable statements about the SECO-WE portfolio can be made because the evaluated projects do not represent the portfolio. • 2023 and 2024 ratings for external project evaluations done by different contractors.
3. What insights do evaluations provide regarding the use and quality of Theories of Change and logframes? To what extent do these elements influence the effectiveness of evaluated projects?	Code system developed by Syspons	Source/Collection: • 48 external project evaluation reports • Interviews (primary data) for triangulation Analysis: • Descriptive statistics: Code frequencies • Quantitative analysis: Crossing of frequencies, and comparison with REQA and DAC assessment data • Qualitative content analysis of coded segments (further differentiation, examples)	• Number of theory-based evaluations and/or such that use a contribution analysis is limited. • Findings will not be generalisable for SECO-WE's wider portfolio.

Evaluation question	Assessment criteria	Data sources, data collection, data analysis	Limitations
4. What evidence do evaluations provide in relation to SECO-WE's impact hypotheses? Where are the strengths and gaps?	• Mapping evaluation findings, REQA and DAC ratings against SECO-WE's impact hypotheses • Focus on impact hypotheses 1.1, 2.2 or 3.1 (formerly 1.4)	Source/Collection: • Ca. 30 project evaluation reports which include information on SECO-WE's impact hypotheses 1.1, 2.2 or 3.1 (1.4) Analysis: • Descriptive statistics: Code frequencies • Quantitative analysis: Crossing of frequencies, and comparison with REQA and DAC assessment data • Qualitative content analysis of coded segments (further differentiation, examples)	• Focus is on 3 out of 8 impact hypotheses only. • Causal analysis relatively rare in the evaluations; hence causal evidence has been limited. • Use of standard indicators to complement the analysis has not been possible due to limited evidence on corresponding outcome levels from evaluation reports.

2 List of evaluations assessed incl. REQA ratings

Table 2: List of Evaluations with respective REQA ratings

Number WEQA	Evaluation Title	Year	Adequacy	Transparency	Triangulation	Appropriateness	Attribution	Contribution	Recommendations
WEHU 211	GPIPR External Mid-Term Evaluation	2023	High	Inadequate	Inadequate	Basic	Not applicable	Not applicable	Basic
WEHU 213	Midterm Evaluation of the ITCMENATEX Programme	2023	High	Inadequate	Basic	Basic	Not applicable	Not applicable	Good
WEHU 214	Sustainable Recycling Industries Phase II (2019-2023)	2023	High	Basic	Basic	Basic	Not applicable	Not applicable	Inadequate
WEHU-WEMU 215	Extractives Global Programmatic Support Trust Fund (EGPS-2) Mid-Term Review	2023	Good	Inadequate	Basic	Basic	Not applicable	Not applicable	Basic
WEIF 131	Final Report on the Evaluation of CIIP for the European Commission	2023	Good	Inadequate	Basic	Inadequate	Not applicable	Not applicable	Basic
WEIF 132	Central Asia Financial Inclusion Project 602131 End-Term Review	2023	High	Inadequate	Basic	Good	Not applicable	Not applicable	Good
WEIF 133	Decentralized External Progress Evaluation of UNDP Project: Strengthening MSME Business Membership Organizations in Ukraine: Phase II (2019-2023)	2023	High	Basic	Basic	Good	Not applicable	Not applicable	Inadequate
WEIF-WEIN 130	Vietnam Country Report Evaluation of the development impact of PIDG	2023	Good	High	Good	High	Not applicable	Good	Good
WEIN 83	End-of-program External Evaluation of the Activities Carried out in Colombia and Peru under the Umbrella of the IFC LAC Sustainable Cities Program	2023	High	Good	Good	Good	Basic	Basic	Good
WEIN 84	External Mid-Term Progress Assessment for Block C Countries Global Water Security & Sanitation Partnership (GWSP)Summary Report: Bangladesh, Ethiopia, Haiti, Pakistan, and Vietnam	2023	Good	Basic	Basic	Basic	Not applicable	Inadequate	Not considered

Number WEQA	Evaluation Title	Year	Adequacy	Transparency	Triangulation	Appropriateness	Attribution	Contribution	Recommendations
WEIN 85	Mid-term Review Hayenna – Integrated Urban Development Project in Egypt (2018-2024)	2023	High	Inadequate	Basic	Inadequate	Not applicable	Not applicable	Basic
WEIN 86	MOLO External Evaluation Report	2023	Good	Not considered	Inadequate	Not considered	Not applicable	Not applicable	Inadequate
WEIN 87	External Interim Evaluation Urban Transformation Project Sarajevo	2023	Good	Not considered	Inadequate	Not considered	Not applicable	Not applicable	Basic
WEIN 88	Utility reform support, TA Fund, Peru	2023	Good	Not considered	Inadequate	Not considered	Not applicable	Not applicable	Inadequate
WEIN 89	Ex-Post-Evaluation of the Technical Assistance to OTASS and the EPS EMAPA San Martin and Moyobamba for the implementation of the Transitional Support Regime	2023	Good	Not considered	Inadequate	Inadequate	Not applicable	Not applicable	Basic
WEMU 109	Azerbaijan Financial Sector Modernization Project (FSMP2), Independent Evaluation	2023	High	Basic	Basic	Basic	Not applicable	Not applicable	Basic
WEMU 110	Nepal Public Financial Management Support Multi Donor Trust Fund (MDTF) Mid-term Evaluation	2023	Not considered	Not considered	Basic	Basic	Not applicable	Not applicable	Basic
WEMU 111	Strengthening Subnational PFM in Albania, External End-of-Phase Evaluation	2023	High	Basic	Basic	Basic	Not applicable	Not applicable	Basic
WEMU 112	Independent Evaluation of the Extractive Industries Transparency Initiative	2023	High	Basic	Good	Good	Not applicable	Basic	Good
WEMU 113	Egmont Group Centre of FIU Excellence and Leadership (ECOFEL)	2023	Good	Good	Good	Good	Not applicable	Not applicable	Good
WEMU 114	Mid-term Evaluation of the Data for Decisions Fund (D4D)	2023	Good	Basic	Basic	Good	Not applicable	Not applicable	Good
WEMU 115	End-of-phase external evaluation of BCC II programme	2023	High	Inadequate	Good	Good	Not applicable	Not applicable	Basic

Number WEQA	Evaluation Title	Year	Adequacy	Transparency	Triangulation	Appropriateness	Attribution	Contribution	Recommendations
WEMU 116	GIZ: Governance for Inclusive Development (GovID)SECO: Domestic Revenue Mobilisation Project Phase III (2016-2023) End of Project Evaluation Report	2023	Good	Basic	Basic	Basic	Not applicable	Not applicable	Inadequate
WEMU 117	PINK – Procurement Infrastructure & Knowledge Management Program	2023	Basic	Inadequate	Inadequate	Inadequate	Not applicable	Not applicable	Basic
WEMU 118	MTR BDT III Independent Mid-term review Swiss Bank Executives' Training Program III (Swiss BET)	2023	Good	Inadequate	Good	Basic	Not applicable	Not applicable	Good
WEQA	Independent evaluation of SECO-WE's climate approach	2023	High	Basic	Basic	Good	Not applicable	Inadequate	Basic
WEHU 216	Independent terminal evaluation Global Quality and Standards Programme (GQSP)	2024	High	Basic	Basic	Good	Not applicable	Good	Good
WEHU 218	Impact Assessment Report: The impacts of GQSP Indonesia SMART-Fish 2 under Global Quality Standard Program (GQSP) on Selected Aquaculture Value Chains in Indonesia	2024	Good	Basic	High	Good	Basic	Not applicable	Inadequate
WEHU 219	Independent Terminal Evaluation of the Global Eco-Industrial Parks Programme (GEIPP)	2024	High	Inadequate	Basic	Inadequate	Not applicable	Basic	Good
WEHU 220	Mid-Term Evaluation of the Swiss Better Gold, Phase III	2024	Good	Basic	Basic	Good	Not applicable	Not applicable	Good
WEHU 222	Productivity Eco-Systems for Decent Work Independent Mid-term Evaluation	2024	Good	Basic	Basic	Basic	Not applicable	Not applicable	Good
WEHU 223	Final Report ITC T4SD Evaluation	2024	Not considered	Not considered	Not considered	Not considered	Not applicable	Not applicable	Not considered
WEHU 224	Third Program Evaluation of the Forest Carbon Partnership Facility	2024	High	Good	Good	High	Not applicable	Good	Basic

Number WEQA	Evaluation Title	Year	Adequacy	Transparency	Triangulation	Appropriateness	Attribution	Contribution	Recommendations
WEHU 225	Impact Evaluation for Global Reporting Initiative (GRI)	2024	Good	Good	Basic	Good	Not applicable	High	Basic
WEHU 226	Navigation Systemic Market Transformation: mid-term review of IDH 2021-2025	2024	High	Good	Good	Good	Not applicable	Basic	Inadequate
WEHU 227	External Evaluation Swiss Trade Policy and Export Promotion Project Vietnam	2024	Basic	Not considered	Inadequate	Inadequate	Not applicable	Not applicable	Inadequate
WEHU-WEIF 221	Evaluation of the Colombia Mas Competitiva Programme - Final Report	2024	High	Good	Good	Good	Basic	Not applicable	Good
WEIF 134	Independent Assessment of Lab Achievements and of Lessons Learned as a Result of SECO Funding from 2019 to 2023	2024	Good	Inadequate	Basic	Basic	Not applicable	Inadequate	Basic
WEIF 135	Impact Evaluation of "Employing Futuro" - First Social Impact Bond in Latin America	2024	Good	High	Good	High	High	Not applicable	Not considered
WEIF 136	Final Evaluation Report EBRD-SECO ITCP Early-Stage Assessment	2024	Good	Basic	Basic	Basic	Not applicable	Basic	Basic
WEIN 91	Ghana Urban Mobility and Accessibility Project (GUMAP) Evaluation Report	2024	Good	Inadequate	Basic	Basic	Not applicable	Not applicable	Basic
WEIN 92	External Mid-term Evaluation and outlook of Renewable Energy Skills Development (RESD) in Indonesia	2024	Good	Good	Good	Good	Not applicable	Basic	Basic
WEIN 93	SURGE Umbrella Mid-Term Evaluation	2024	High	Basic	Basic	Basic	Not applicable	Basic	Good
WEIN 94	External evaluation of the African Cities Lab in urban development project	2024	High	Basic	Good	Basic	Not applicable	Inadequate	Basic
WEMU 119	Macroeconomic Planning and Management Project (MPMP) End of Phase Evaluation	2024	High	Inadequate	Basic	Good	Not applicable	Good	Basic
WEMU 120	Program Evaluation and Impact Assessment of the Global Program for Sustainability: Mid-Term Review Report	2024	High	Inadequate	Basic	Good	Not applicable	Good	Basic

Number WEQA	Evaluation Title	Year	Adequacy	Transparency	Triangulation	Appropriateness	Attribution	Contribution	Recommendations
WEMU 121	Revenue Administration and Public Financial Management Reform in Southeast Europe (SEE II) project mid-term external evaluation	2024	High	Basic	Basic	Good	Not applicable	Not applicable	Good
WEMU 122	Etude Bilan Prospective Du Projet D'Assistance Technique et Financiere a la Direction General des Impots (PATF/DGI) du Burkina Faso	2024	Good	Inadequate	Basic	Good	Not applicable	Basic	High

3 List of evaluations assessed incl. DAC criteria ratings

Table 3: List of Evaluations with respective OECD-DAC ratings

Number WEQA	Evaluation Title	Year	Relevance	Coherence	Effectiveness	Efficiency	Impact	Sustainability
WEHU 211	GPIPR External Mid-Term Evaluation	2023	Highly satisfactory	Satisfactory	Satisfactory	Satisfactory	Highly satisfactory	Highly satisfactory
WEHU 213	Midterm Evaluation of the ITCMEN-ATEX Programme	2023	Satisfactory	Unsatisfactory	Satisfactory	Unsatisfactory	Satisfactory	Satisfactory
WEHU 214	Sustainable Recycling Industries Phase II (2019-2023)	2023	Highly satisfactory	Satisfactory	Satisfactory	Satisfactory	Highly satisfactory	Satisfactory
WEHU-WEMU 215	Extractives Global Programmatic Support Trust Fund (EGPS-2) Mid-Term Review	2023	Highly satisfactory	Not assessed	Satisfactory	Satisfactory	Satisfactory	Unsatisfactory
WEIF 131	Final Report on the Evaluation of CIIP for the European Commission	2023	Satisfactory	Unsatisfactory	Unsatisfactory	Satisfactory	Unsatisfactory	Satisfactory
WEIF 132	Central Asia Financial Inclusion Project 602131 End-Term Review	2023	Satisfactory	Satisfactory	Satisfactory	Satisfactory	Not assessed	Satisfactory
WEIF 133	Decentralized External Progress Evaluation of UNDP Project:	2023	Satisfactory	Highly satisfactory	Satisfactory	Satisfactory	Satisfactory	Unsatisfactory

Number WEQA	Evaluation Title	Year	Relevance	Coherence	Effectiveness	Efficiency	Impact	Sustainability
	Strengthening MSME Business Membership Organizations in Ukraine: Phase II (2019-2023)							
WEIF-WEIN 130	Vietnam Country Report Evaluation of the development impact of PIDG	2023	Highly satisfactory	Unsatisfactory	Highly satisfactory	Satisfactory	Satisfactory	Not assessed
WEIN 83	End-of-program External Evaluation of the Activities Carried out in Colombia and Peru under the Umbrella of the IFC LAC Sustainable Cities Program	2023	Satisfactory	Highly satisfactory	Unsatisfactory	Satisfactory	Satisfactory	Satisfactory
WEIN 84	External Mid-Term Progress Assessment for Block C Countries Global Water Security & Sanitation Partnership (GWSP)Summary Report: Bangladesh, Ethiopia, Haiti, Pakistan, and Vietnam	2023	Highly satisfactory	Satisfactory	Satisfactory	Not assessed	Not assessed	Not assessed
WEIN 85	Mid-term Review Hayenna – Integrated Urban Development Project in Egypt (2018-2024)	2023	Satisfactory	Satisfactory	Unsatisfactory	Satisfactory	Satisfactory	Satisfactory
WEIN 86	MOLO External Evaluation Report	2023	Highly satisfactory	Highly satisfactory	Satisfactory	Satisfactory	Not assessed	Satisfactory
WEIN 87	External Interim Evaluation Urban Transformation Project Sarajevo	2023	Unsatisfactory	Not assessed	Satisfactory	Unsatisfactory	Not assessed	Not assessed
WEIN 88	Utility reform support, TA Fund, Peru	2023	Highly satisfactory	Not assessed	Satisfactory	Satisfactory	Satisfactory	Satisfactory
WEIN 89	Ex-Post-Evaluation of the Technical Assistance to OTASS and the EPS EMAPA San Martin and Moyobamba for the implementation of the Transitional Support Regime	2023	Highly satisfactory	Highly satisfactory	Highly satisfactory	Highly satisfactory	Satisfactory	Satisfactory
WEMU 109	Azerbaijan Financial Sector Modernization Project (FSMP2), Independent Evaluation	2023	Highly satisfactory	Satisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory

Number WEQA	Evaluation Title	Year	Relevance	Coherence	Effectiveness	Efficiency	Impact	Sustainability
WEMU 110	Nepal Public Financial Management Support Multi Donor Trust Fund (MDTF) Mid-term Evaluation	2023	Satisfactory	Satisfactory	Satisfactory	Unsatisfactory	Not assessed	Not assessed
WEMU 111	Strengthening Subnational PFM in Albania, External End-of-Phase Evaluation	2023	Satisfactory	Satisfactory	Satisfactory	Satisfactory	Not assessed	Satisfactory
WEMU 112	Independent Evaluation of the Extractive Industries Transparency Initiative	2023	Satisfactory	Not assessed	Satisfactory	Satisfactory	Satisfactory	Satisfactory
WEMU 113	Egmont Group Centre of FIU Excellence and Leadership (ECOFEL)	2023	Highly satisfactory	Highly satisfactory	Highly satisfactory	Unsatisfactory	Not assessed	Satisfactory
WEMU 114	Mid-term Evaluation of the Data for Decisions Fund (D4D)	2023	Satisfactory	Satisfactory	Satisfactory	Satisfactory	Satisfactory	Unsatisfactory
WEMU 115	End-of-phase external evaluation of BCC II programme	2023	Highly satisfactory	Highly satisfactory	Satisfactory	Satisfactory	Satisfactory	Satisfactory
WEMU 116	GIZ: Governance for Inclusive Development (GovID) SECO: Domestic Revenue Mobilisation Project Phase III (2016-2023) End of Project Evaluation Report	2023	Satisfactory	Highly satisfactory	Satisfactory	Highly satisfactory	Highly satisfactory	Satisfactory
WEMU 117	PINK – Procurement Infrastructure & Knowledge Management Program	2023	Satisfactory	Unsatisfactory	Satisfactory	Unsatisfactory	Not assessed	Unsatisfactory
WEMU 118	MTR BDT III Independent Mid-term review Swiss Bank Executives' Training Program III (Swiss BET)	2023	Satisfactory	Not assessed	Satisfactory	Satisfactory	Not assessed	Satisfactory
WEHU 216	Independent terminal evaluation Global Quality and Standards Programme (GQSP)	2024	Satisfactory	Satisfactory	Unsatisfactory	Satisfactory	Satisfactory	Unsatisfactory
WEHU 218	Impact Assessment Report: The impacts of GQSP Indonesia SMART-Fish 2 under Global Quality Standard	2024	Not assessed	Not assessed	Satisfactory	Not assessed	Highly satisfactory	Not assessed

Number WEQA	Evaluation Title	Year	Relevance	Coherence	Effectiveness	Efficiency	Impact	Sustainability
	Program (GQSP) on Selected Aquaculture Value Chains in Indonesia							
WEHU 219	Independent Terminal Evaluation of the Global Eco-Industrial Parks Programme (GEIPP)	2024	Satisfactory	Satisfactory	Unsatisfactory	Satisfactory	Unsatisfactory	Highly unsatisfactory
WEHU 220	Mid-Term Evaluation of the Swiss Better Gold, Phase III	2024	Satisfactory	Satisfactory	Unsatisfactory	Satisfactory	Not assessed	Unsatisfactory
WEHU 222	Productivity Eco-Systems for Decent Work Independent Mid-term Evaluation	2024	Highly satisfactory	Satisfactory	Unsatisfactory	Satisfactory	Satisfactory	Satisfactory
WEHU 223	Final Report ITC T4SD Evaluation	2024	Satisfactory	Not assessed	Not assessed	Not assessed	Not assessed	Not assessed
WEHU 224	Third Program Evaluation of the Forest Carbon Partnership Facility	2024	Highly satisfactory	Satisfactory	Satisfactory	Satisfactory	Satisfactory	Satisfactory
WEHU 225	Impact Evaluation for Global Reporting Initiative (GRI)	2024	Not assessed	Not assessed	Not assessed	Not assessed	Satisfactory	Not assessed
WEHU 226	Navigating Systemic Market Transformation: mid-term review of IDH 2021-2025	2024	Satisfactory	Highly satisfactory	Satisfactory	Satisfactory	Satisfactory	Satisfactory
WEHU 227	External Evaluation Swiss Trade Policy and Export Promotion Project Vietnam	2024	Satisfactory	Highly satisfactory	Highly unsatisfactory	Highly unsatisfactory	Not assessed	Unsatisfactory
WEHU-WEIF 221	Evaluation of the Colombia Mas Competitiva Programme - Final Report	2024	Satisfactory	Highly unsatisfactory	Satisfactory	Unsatisfactory	Satisfactory	Unsatisfactory
WEIF 134	Independent Assessment of Lab Achievements and of Lessons Learned as a Result of SECO Funding from 2019 to 2023	2024	Highly satisfactory	Satisfactory	Satisfactory	Satisfactory	Satisfactory	Satisfactory

Number WEQA	Evaluation Title	Year	Relevance	Coherence	Effectiveness	Efficiency	Impact	Sustainability
WEIF 135	Impact Evaluation of "Employing Futuro" - First Social Impact Bond in Latin America	2024	Not assessed	Not assessed	Not assessed	Not assessed	Highly satisfactory	Not assessed
WEIF 136	Final Evaluation Report EBRD-SECO ITCP Early-Stage Assessment	2024	Satisfactory	Satisfactory	Highly satisfactory	Satisfactory	Satisfactory	Satisfactory
WEIN 91	Ghana Urban Mobility and Accessibility Project (GUMAP) Evaluation Report	2024	Unsatisfactory	Satisfactory	Unsatisfactory	Not assessed	Unsatisfactory	Unsatisfactory
WEIN 92	External Mid-term Evaluation and outlook of Renewable Energy Skills Development (RESD) in Indonesia	2024	Highly satisfactory	Highly satisfactory	Satisfactory	Highly satisfactory	Satisfactory	Satisfactory
WEIN 93	SURGE Umbrella Mid-Term Evaluation	2024	Highly satisfactory	Highly satisfactory	Satisfactory	Satisfactory	Not assessed	Unsatisfactory
WEIN 94	External evaluation of the African Cities Lab in urban development project	2024	Satisfactory	Satisfactory	Satisfactory	Unsatisfactory	Not assessed	Highly unsatisfactory
WEMU 119	Macroeconomic Planning and Management Project (MPMP) End of Phase Evaluation	2024	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory	Not assessed	Unsatisfactory
WEMU 120	Program Evaluation and Impact Assessment of the Global Program for Sustainability: Mid-Term Review Report	2024	Highly satisfactory	Not assessed	Satisfactory	Satisfactory	Satisfactory	Unsatisfactory
WEMU 121	Revenue Administration and Public Financial Management Reform in Southeast Europe (SEE II) project mid-term external evaluation	2024	Highly satisfactory	Highly satisfactory	Unsatisfactory	Satisfactory	Satisfactory	Unsatisfactory
WEMU 122	Etude Bilan Prospective Du Projet D'Assistance Technique et Financiere a la Direction General des Impots (PATF/DGI) du Burkina Faso	2024	Satisfactory	Highly satisfactory	Satisfactory	Satisfactory	Satisfactory	Satisfactory

4 Detailed methodological approach: Rapid Evidence Quality Assessment (REQA)

Detailed description of scoring system

For the Rapid Evidence Quality Assessment (REQA), evaluation reports are assessed along the following five criteria: 1) Adequacy of the evaluation for its intended purpose, 2) Transparency of methods, 3) Triangulation of data and perspectives, 4) Appropriateness of methods, 5) Strength of causal analysis - 5.1) Strength of attribution analysis or 5.2) Strength of analysis of contribution - and 6) Usability of recommendations.

Each quality criterion is assessed according to the following levels (unbalanced scale):

0 - not considered	The criterion was not addressed, or insufficient information is available. This rubric level indicates that an evaluation report provides no indication that the criterion was taken into consideration. At this level, a score of 0 reflects a deficiency, as it indicates that the criterion was applicable but not addressed or supported by sufficient information.
1 - inadequate	The criterion was addressed, however to an insufficient extent.
2 - basic	The criterion was addressed in a basic manner.
3 - good	The criterion was addressed well.
4 – high	The criterion was addressed very well, and the evaluation / evidence product is of high quality for this criterion.

Each level is accompanied by specific descriptors (description of what a level looks like for each criterion). Further, for some of the criteria, sub-criteria with descriptors are developed to differentiate between key dimensions within these criteria.

1. Adequacy of the evaluation for its intended purpose

This criterion assesses whether evaluation[33] is an appropriate assessment tool for the purpose and objectives at hand. This criterion does not consider the specific evaluation approach, design or methods but is based on the evaluation objective and purpose (as stated in the evaluation report). The extent to which the methods are appropriate is considered in criterion 4.

0 - not considered	*1 - inadequate*	*2 - basic*	*3 - good*	*4 – high*
No clear evaluation purpose or objective is stated, or the assessment is not recognisable as an evaluation.	The purpose and objectives are stated, but another assessment tool would be more suitable to meet them. Evaluation appears misapplied.	Evaluation is one of several tools that could partially meet the purpose and objectives, but its added value is not clearly established.	Evaluation is a mostly suitable tool to address the stated purpose and objectives. Its role is generally justified and appropriate.	Evaluation is clearly the appropriate tool for the stated purpose and objectives. Its use is well-justified and aligned with what evaluation can meaningfully deliver.

[33] Evaluation are characterized by: (i) An evaluation is the systematic and objective assessment of a (or several) planned, on-going or completed project or programme (or any of its components), covering its design, implementation, and/or results. (ii) Through evaluation, SDC and SECO-WE aim to determine relevance, coherence, effectiveness, efficiency, impact, and/or sustainability. To ensure accuracy and feasibility, evaluations should focus on the most relevant criteria, maintaining clarity of purpose and objectives. (iii) Evaluations should be useful, credible, feasible, accurate, independent, and correct, helping to incorporate learnings into decision-making by SDC or SECO, their partners, and recipients.

2. Transparency of methods

Transparency of methods assesses whether an evaluation clearly describes and justifies i) sample and ii) data collection (incl. sampling strategy) and iii) analysis methods. The levels are further based on whether the evaluation identifies and discusses related limitations (incl. biases). Note that this assessment focuses on the transparency of methods, not their appropriateness. Therefore, the criterion is to judge whether a description is available and clear, not to assess the content – this will be covered in the following criteria. This assessment is further based on the extent to which the relevant information is easily accessible for the reader, in most cases the assessment will therefore be based on a methodological section or other dedicated parts of the evaluation report (for example subsections, tables or methodological annexes). If the information is scattered or needs to be inferred from the analysis or other parts of the report, the transparency of methods is considered low. Overall, the criterion is divided into three sub-criteria for each of which the level is assessed. The overall level is typically based on a weighted average of the sub-criteria as indicated in the table.

Sub-criteria	*0 - not considered*	*1 - inadequate*	*2 - basic*	*3 - good*	*4 – high*
Sample *(33%)*	The evaluation does not include parts of the report dedicated to the sample.	Sample size and composition are not described.	Sample size and composition are described.	Sample size and composition are described and justified.	Sample size, and composition, are described and justified, and limitations are discussed.
Data collection methods *(33%)*	The evaluation does not include parts of the report dedicated to describing the data collection methods.	Data collection methods (incl. sampling strategy) are not described.	Data collection methods (incl. sampling strategy) as well as some limitations are outlined.	Data collection methods (incl. sampling strategy) are well described, and limitations explained.	Data collection methods (incl. sampling strategy) are well described and justified taking limitations into account.
Data analysis methods *(33%)*	The evaluation does not include parts of the report dedicated to describing data analysis methods.	Data analysis methods are unclear.	Data analysis methods as well as some limitations are outlined.	Data analysis methods are well described, and limitations explained.	Data analysis methods are well described and justified taking limitations into account.

3. Triangulation of data and perspectives

Triangulation assesses whether i) multiple methods of data collection are used[34], ii) different stakeholder perspectives considered and iii) conflicting findings (methods, stakeholders, researchers) discussed to enhance the credibility and rigour of findings and conclusions. For this purpose, the criterion is divided into three sub-criteria for each of which the level is assessed. The overall level is typically based on a weighted average of the sub-criteria as indicated in the table.

Sub-criterion	*0 - not considered*	*1 – inadequate*	*2 - basic*	*3 - good*	*4 – high*
Data collection methods *(40%)*	Data collection methods used are not clearly stated.	The evaluation relies on a single data collection method.	The evaluation uses one data collection method but references other findings/sources.	The evaluation employs two or more data collection methods.	The evaluation uses complementary and distinct data collection methods -- including both quantitative and qualitative collection methods.
Perspectives *(30%)*	Data sources/ perspectives are not clearly stated.	The evaluation lacks multiple perspectives.	The evaluation presents multiple perspectives.	The evaluation discusses multiple perspectives.	The evaluation discusses all relevant perspectives thoroughly.
Conflicting finding *(30%)*	-	The evaluation does not present conflicting or nuanced findings.	The evaluation acknowledges conflicting or nuanced findings without exploration.	The evaluation explores conflicting or nuanced findings.	The evaluation explores conflicting and nuanced findings and their implications.

4. Appropriateness of methods

Appropriateness assesses whether the evaluation's i) approach, design and methods, ii) data and sample, iii) and analysis are suitable for the evaluation's (or evidence product's) purpose and objectives and can generate reliable and valid findings and conclusions. Based on this, the criterion is divided into three sub-criteria for each of which the level is assessed. The overall level is typically based on a weighted average of the sub-criteria as indicated in the table.

Sub-criteria	*0 - not considered*	*1 - inadequate*	*2 - basic*	*3 - good*	*4 – high*
Methods *(33%)*	The evaluation does not include enough information for allowing to judge appropriateness.	The evaluation approach, design and methods are not relevant to the purpose and objectives.	The evaluation approach, design and methods are partly relevant.	The evaluation approach, design and methods are mostly relevant.	The evaluation approach, design and methods are highly relevant.

[34] Different data collection methods can include: i) interviews (in-depth and key informant), ii) focus group discussions, iii) surveys or iv) other. Document review would be expected for levels 2 and above and is only counted as distinct methods when the review is systematic and extensive.

Data *(33%)*	The evaluation does not include enough information for allowing to judge appropriateness.	The data is unreliable or the sample (size and composition) inadequate for findings.	The data has reliability concerns, or the sample (size and composition) only allow for limited findings.	The data is reliable, and the sample (size and composition) is adequate.	The data is highly reliable and includes quality controls, and the sample (size and composition) is adequate.
Data analysis *(33%)*	The evaluation does not include enough information for allowing to judge appropriateness	The data analysis is unclear or not stringent, in particular, findings and/or conclusions are only partially linked to the data with substantial (logical) gaps.	The data analysis is not always stringent, in particular, findings and/or conclusions are only partially linked to the data with substantial (logical) gaps.	The data analysis is systematic and convincing, in particular, findings and/or conclusions are linked to the reported data.	The data analysis is systematic and highly stringent, in particular, findings and/or conclusions are clearly derived from data with limitations acknowledged.

5. Strength of causal analysis

Strength of causal analysis assesses whether an evaluation effectively identifies and justifies the intervention's role in producing observed changes, considers a control or comparison group or the influence of external factors and/or alternative explanations. This criterion is applicable only if an evaluation assesses causal questions. Questions related to a project or program's effectiveness and impact (and sustainability of effects) are usually causal in nature and causal claims are expected in end-term and ex-post evaluations.

Depending on the nature of causal analysis which has been performed (as stated in the evaluation report), criterion 5a (attribution) and/or 5b (contribution) is used. If the evaluation is categorized as 'impact evaluation (attribution)' or 'impact evaluation (contribution)', the respective criterion is to be used, if it is dedicated 'mixed' or 'process evaluation', both 5a and 5b, one or neither is to be used, depending on the evaluation.

5.1. Strength of attribution analysis

Causal attribution assesses whether an evaluation establishes a direct causal link between an intervention and observed changes, using an appropriate counterfactual design with control or comparison data.

0 – not applicable	*1 - inadequate*	*2 - basic*	*3 - good*	*4 – high*
The evaluation objective covers causal claims but does not consider causal analysis, making it impossible to judge the criterion. Or there is insufficient information to judge the criterion.	No control/ comparison data is used.	Control/ comparison data is used, but without clear justification; causal links between intervention and outcomes are explored superficially.	Control/ comparison data is used with reflection of bias; causal links and underlying assumptions are analysed; alternative explanations are considered.	High-quality control/ comparison data with strong justification and no or minimal bias; causal links and assumptions are rigorously tested; the intervention's specific effect is isolated.

5.2. Strength of analysis of contribution

Causal contribution assesses whether an evaluation explains how, why or under what configuration an intervention contributed to observed changes (following a mechanism-based/generative or configurational design), considering the influence of external factors/conditions and/or alternative explanations.

0 – not applicable	*1 - inadequate*	*2 - basic*	*3 - good*	*4 – high*
The evaluation objective covers causal claims but does not consider causal analysis, making it impossible to judge the criterion. Or there is insufficient information to judge the criterion.	The intervention's role in producing observed outcomes is not explored; (alternative) contributing (external) factors/conditions are ignored.	The intervention's contribution is acknowledged, but the analysis is weak; (alternative) contributing (external) factor/conditions are referenced but not analysed.	The intervention's contribution is clearly analysed, with consideration of (alternative) contributing (external) factors/conditions.	The intervention's contribution is comprehensively analysed, by using evidence tests, and discussing the relative influence of (alternative) contributing (external) factors/conditions.

6. Useability of recommendations

The criterion for recommendations does not capture the inherent evidence value and is divided into 4 sub-criteria. For each sub-criteria the level is assessed. The overall level is typically based on a weighted average of the sub-criteria as indicated in the table. These sub-criteria are mostly based on the GPK recommendation and assess whether recommendations in the evaluation are i) actionable[14], ii) actor addressed[15], and iii) prioritised[16]. Further, the criterion assesses if iv) recommendations are logically linked to findings and conclusions. Similarly to the other sub-criteria, this link is assessed based on formal consistency – i.e. whether a logical connection is included – rather than on the appropriateness or quality of the recommendation itself. The different levels of the sub-criteria primarily distinguish between the quantity of recommendations which fulfil the sub-criteria rather than the quality with which the sub-criteria are met. A good qualitative assessment of the extent to which recommendations are for example actionable or well prioritised would require subject-matter expertise and thus be less reliable and not fit for the purpose of this rapid assessment.

Sub-criteria	*0 - not considered*	*1 - inadequate*	*2 - basic*	*3 - good*	*4 – high*
Actionable *(20%)*	The evaluation does not include clearly presented recommendations.	Recommendations are not actionable.	Recommendations are partly actionable.	Recommendations are largely actionable.	All recommendations are actionable.
Actor-addressed *(20%)*		Recommendations are not actor-addressed.	Recommendations are partly actor-addressed.	Recommendations are largely actor-addressed.	All recommendations are actor-addressed.

Prioritised *(20%)*		No prioritisation is provided, even where the number of recommendations would clearly require it.	Some recommendations are prioritised, but the approach is insufficient, particularly when recommendations are numerous.	Most recommendations are prioritised appropriately, with clearer structuring when they are more numerous.	All recommendations are clearly prioritised, with the level of prioritisation suited to their number.
Linked to findings *(40%)*		Recommendations are not linked to findings.	Recommendations are partly linked to findings.	Recommendations are largely linked to findings.	All recommendations are linked to findings.

Assessment process

The assessment of the 49 evaluation reports from 2023 and 2024 was conducted through a **two-step review process**, complemented by an additional quality assurance of the results. In an initial review, each of the evaluation reports was reviewed by a consultant from Syspons. The main text of the report as well as any available annexes, including in separate documents, were considered during the review. Inception reports, Terms of Reference of evaluations, or comparable documents were not taken into account. In addition to the rating along the six REQA criteria with the respective sub-criteria (see above), Syspons extracted and classified further information such as the type of the evaluation, approach and data collection methods used. Each sub-criterion was rated on a 5-point scale from 0 (not considered) to 4 (high quality). In addition to the rating, a brief justification of 1-2 sentences was provided referring to 1-3 passages in the evaluation report exemplifying the reasoning behind the rating. During the initial review process any marginal ratings or difficult-to-rate cases were highlighted for the quality assurance. After the initial review, each evaluation report was reviewed by a different consultant from Syspons with a focus on the highlighted aspects from the initial review as well as consistency between rating and justification.

5 Detailed methodological approach: OECD-DAC Assessment

Detailed description of scoring system

The assessment of project performance is based on the OECD-DAC evaluation criteria, using a structured rating methodology developed by SECO-WE and SDC. This standardised approach enables consistent, transparent, and quantifiable assessments of project results across six DAC criteria: relevance, coherence, effectiveness, efficiency, impact, and sustainability. Each of the six DAC criteria is broken down into multiple sub-criteria, typically three to five per criterion.

Each OECD-DAC sub-criterion is assessed according to the following levels:

	Relevance / coherence / efficiency	*Effectiveness*	*Impact*	*Sustainability*
1= Highly satisfactory	There were no shortcomings in relation to the intervention's relevance/ coherence/ efficiency.	Objectives at outcome level were (or are likely to be) fully achieved or exceeded.	The intervention had (or is likely to have) a significant positive impact.	All of the intervention's benefits (will) last. *Note: for this rating, clear evidence is required (not only assumptions).*
2= Satisfactory	There were moderate shortcomings in relation to the intervention's relevance/ coherence/ efficiency.	Objectives at outcome level were (or are likely to be) largely achieved.	The intervention had (or is likely to have) an overall positive impact.	A majority of the intervention's benefits (will) last.
3= Unsatisfactory	There were important shortcomings in relation to the intervention's relevance/ coherence/ efficiency.	Objectives at outcome level were (or are likely to be) only partially achieved (at a rather low level). *Note: if outputs are achieved, but do not result in the expected outcomes, consider rating effectiveness as unsatisfactory.*	The intervention had (or is likely to have) no impact.	A minority of the intervention's benefits (will) last.
4= Highly unsatisfactory	There were very severe shortcomings in relation to the intervention's relevance/ coherence/ efficiency.	Objectives at outcome level were not achieved (or are unlikely to be achieved).	The intervention had (or is likely to have) an unexpected negative impact.	None of the intervention's benefits (will) last.
0= Not assessed	The criteria statement cannot be assessed. Please explain in the justifications section.			

For the assessment, Syspons further developed specific descriptors (description of what a level looks like for each sub-criterion).

1. Relevance: Is the intervention doing the right things?

The extent to which the intervention's objectives and design (at the time of design and at time of evaluation) respond to beneficiaries' and involved stakeholders' needs and priorities and continue to do so if circumstances change.

Note: Understanding gendered power dynamics and reflecting on the SDG commitment to "leave no one behind" are crucial in understanding relevance.

Sub-criteria	*1 – Highly satisfactory*	*2 - Satisfactory*	*3 - Unsatisfactory*	*4 – Highly unsatisfactory*
Responsiveness to needs, policies and priorities	Fully aligned with beneficiary and stakeholder needs. No trade-offs, or they are clearly explained and well managed.	Generally aligned, but moderate gaps or unclear trade-offs noted in the report.	Important needs or priorities overlooked. Report highlights significant misalignment or weak trade-off management.	Major misalignment. Needs and priorities not reflected; no clear needs analysis provided.
Sensitiveness and responsiveness to the context and capacities of the beneficiaries and involved stakeholders	Contextual and capacity factors clearly considered in design based on evidence provided.	Most relevant factors considered, but some gaps or limited depth.	Key context or capacity aspects not sufficiently addressed. Design shows important oversights.	Context and capacities ignored or seriously misjudged. Design disconnected from local realities.
Quality of design	Design is coherent and feasible. Objectives, indicators, and exit strategy are clear, realistic, and well aligned with needs.	Design is generally sound but with moderate issues (e.g., unclear indicators, weak exit planning).	Design has major flaws (e.g., vague logic, unrealistic objectives, missing sustainability elements).	Design is fundamentally flawed or absent. Serious issues undermine intervention relevance.
Adaptation over time	Strong, timely, and effective adaptation to relevant contextual changes is clearly documented.	Adaptation occurred but was delayed or only partially effective.	Limited adaptation. Key changes were poorly addressed or recognized late.	No meaningful adaptation despite significant contextual shifts. Intervention remained rigid.

2. Coherence: How well does the intervention fit?

The compatibility of the evaluated intervention with other interventions in a country, sector or institution, i.e., the extent to which other interventions (in particular policies) support or undermine the intervention and vice versa.

Sub-criteria	*1 – Highly satisfactory*	*2 - Satisfactory*	*3 - Unsatisfactory*	*4 – Highly unsatisfactory*
Internal policy alignment	Strong alignment with Swiss development strategies (global, regional, country level) and relevant international norms/agreements. Clearly evidenced.	Generally aligned, but report notes moderate inconsistencies or areas where alignment is less clearly articulated.	Important misalignments or omissions noted. Links to Swiss or international frameworks are weak or partially addressed.	Little or no alignment with Swiss policy frameworks or international norms. Serious inconsistencies or a lack of reference in design or implementation.
Internal com-	Strong evidence of coherence with other Swiss interventions. Synergies,	Generally compatible, with some minor overlaps or	Several important overlaps or gaps. Report shows lim-	Intervention is isolated or duplicative. No evidence of co-

patibility	complementarities, and coordination clearly described. No duplication.	missed opportunities for coordination. Complementarity is noted but not fully realized.	ited coordination with related Swiss efforts or other parts of the Swiss government.	ordination; inconsistencies with other Swiss efforts in the same context.
External compatibility	Intervention is well-coordinated with other actors. Strong complementarity, use of local systems, and value-added clearly documented.	Some coordination or complementarity with other actors is evident, but moderate gaps or overlaps are noted.	Important compatibility issues identified. The intervention duplicates efforts or fails to coordinate with key actors.	o meaningful coordination. Intervention operates in isolation and may undermine or contradict efforts of other actors.

3. Effectiveness: Is the intervention achieving its objectives?

The extent to which the intervention achieved, or is expected to achieve, its objectives and its results, including any differential results across groups.

Sub-criteria	*1 – Highly satisfactory*	*2 - Satisfactory*	*3 - Unsatisfactory*	*4 – Highly unsatisfactory*
Achievement of objectives	Most or all outcome-level objectives, including key transversal ones, fully achieved or exceeded. Strong supporting evidence.	Most outcome-level objectives largely achieved. Some shortfalls noted, but no critical gaps in core objectives.	Important misalignments or omissions noted. Links to Swiss or international frameworks are weak or partially addressed.	Objectives not achieved or unlikely to be. No meaningful outcome-level progress reported despite output delivery.
Unintended effects	Unintended effects (both positive and negative) are clearly identified. Intervention responded appropriately, managing risks and leveraging opportunities.	Some unintended effects noted. Response was adequate, though could have been more proactive or comprehensive.	Important unintended effects were insufficiently addressed or only acknowledged after the fact. Response was reactive or partial.	Unintended effects (especially negative ones) were ignored, denied, or poorly managed. Evaluation identifies serious blind spots.
Differential results	Clear evidence that results were inclusive and equitable. Intervention actively applied key principles (e.g., gender equity, accountability, participation).	Principles of inclusion and equity were generally applied, but with some unevenness or moderate gaps across groups.	Inclusion was weak. Some groups were left behind or not sufficiently reached. Limited application of key equity principles.	Results were inequitable or exclusionary. Little or no attention to inclusion, participation, or non-discrimination.

4. Efficiency: How well are resources being used?

The extent to which the intervention delivers, or is likely to deliver, results in an economic and timely way.

Sub-criteria	*1 – Highly satisfactory*	*2 - Satisfactory*	*3 - Unsatisfactory*	*4 – Highly unsatisfactory*
Economic efficiency	Resources were used economically to achieve results. Cost-efficiency was actively pursued and achieved.	Generally cost-efficient with some moderate inefficiencies or missed opportunities to optimize resource use.	Notable inefficiencies in resource use. Report identifies important cost issues or suboptimal allocation between target groups or time periods.	Serious inefficiencies. High costs, poor resource use, or lack of cost-awareness undermined value for money.
Timeliness	Outputs/outcomes delivered on or ahead of schedule. Any minor delays were well managed and had no significant impact.	Some delays occurred but were moderate and reasonably mitigated. Timelines mostly respected.	Important delays affected implementation. Mitigation measures were weak or insufficient.	Severe delays with no or poor mitigation. Timeliness seriously undermined intervention performance, even if external factors contributed.
Operational efficiency	Strong planning, risk management, monitoring, and adaptive steering ensured efficient implementation. Systems worked well.	Mechanisms were generally effective, but some moderate weaknesses (e.g., late budget reallocation, monitoring gaps).	Management or oversight problems led to inefficiencies. Key systems (e.g., procurement, spending, risk control) underperformed.	Major operational failures. Management and monitoring systems were inadequate, contributing to inefficiencies.

5. Impact: What difference does the intervention make?

The extent to which the intervention has generated or is expected to generate significant positive or negative, intended or unintended, higher-level effects. Impact addresses the ultimate significance and potentially transformative effects of the intervention. It seeks to identify social, environmental and economic indirect, secondary and potential consequences of the intervention that are longer term or broader in scope than those already captured under the effectiveness criterion. It does so by examining the holistic and enduring changes in systems or norms, and potential effects on people's well-being, human rights, gender equality, and the environment.

Note: depending on the timing of the evaluation and the timescale of intended benefits, evaluators can assess for both actual impacts (i.e., already evident) and foreseeable impacts.

Sub-criteria	*1 – Highly satisfactory*	*2 - Satisfactory*	*3 - Unsatisfactory*	*4 – Highly unsatisfactory*
Intended impacts	Significant positive long-term changes clearly achieved or likely. Strong evidence for (likely/plausible) systemic or transformational change.	Overall positive changes observed or expected. Some higher-level effects achieved, though more limited in scope or depth.	No meaningful long-term impact achieved or expected. Higher-level objectives largely unmet.	Intervention had or is likely to have negative or regressive effects at the systemic or long-term level.
Contribution to in-	Clear and credible evidence of significant contribution to the intended	Contribution to impact is plausible and moderately evid-	Contribution is weak, unclear, or marginal. Impact	Intervention had no contribution or a negative contribu-

tended impacts	impacts. Role of the intervention is well demonstrated.	enced, though some uncertainty remains.	claims are not well supported by evidence.	tion to intended impacts. Impact logic not credible.
Unintended impacts	Positive unintended impacts occurred and were meaningful. Intervention adapted or leveraged these impacts effectively.	Mostly positive or neutral unintended impacts. Some value-added, with limited risk.	Negative or risky unintended impacts occurred. Mitigation was partial or delayed.	Serious negative unintended impacts (e.g., harm to groups, escalation of tensions, exclusion). Little or no mitigation.
Differential impact	Impacts clearly benefited all key groups. Strong evidence of inclusive and equitable distribution of benefits.	Impacts were broadly inclusive, with moderate variation across groups. Some principles applied, though not consistently.	Noticeable gaps in inclusion. Some groups were disadvantaged or less reached. Principles not systematically applied.	Impacts were inequitable or exclusionary. Some groups negatively affected or left behind.

6. Sustainability: Will the benefits last?

The extent to which the net benefits of the intervention continue or are likely to continue. Includes an examination of the enabling environment for sustainable development, i.e., financial, economic, social, environmental, and institutional capacities of the systems needed to sustain net benefits over time. Involves analysis of resilience, risks and potential trade-offs.

Note: depending on the timing of the evaluation and the timescale of intended benefits, evaluators can assess for both actual sustainability (i.e., the continuation of net benefits created by the intervention that are already evident) and prospective sustainability (i.e., the net benefits for key stakeholders that are likely to continue into the future)

Sub-criteria	*1 – Highly satisfactory*	*2 - Satisfactory*	*3 - Unsatisfactory*	*4 – Highly unsatisfactory*
Capacity and resilience development	Stakeholders (e.g., institutions, communities) have strengthened relevant capacities and resilience. Ownership and political will are clearly established.	Moderate capacity strengthening and resilience-building. Some gaps remain, but ownership is present.	Limited capacity or resilience gains. Risk of backsliding is significant. Ownership is weak or uneven.	No significant capacity gains or resilience. Results depend on external actors or short-term measures.
Financial sustainability	Adequate financial resources (e.g., national budget, partner funds) are secured for continuation.	Some financial commitments are in place. Sustainability is plausible, though not fully secured.	Funding is uncertain or insufficient. Only parts of the intervention can be sustained financially.	No financial basis exists to sustain benefits. High risk of complete discontinuation.
Contextual factors	Strong alignment with reforms or systems that support long-term sustainability. Enabling environment is clearly in place.	Context supports sustainability to some extent. Some reforms or systems are in progress or partially aligned.	Sustainability is hindered by policy, institutional, or governance gaps. Reform efforts are weak or stalled.	Context actively undermines sustainability (e.g., policy reversal, conflict, poor governance). No enabling environment.

Assessment process

The overall score for each DAC criterion is calculated as the unweighted mean of the ratings across its respective sub-criteria. A narrative justification is provided for each score, summarising the underlying evidence and rationale. 23 evaluation reports from 2024 were rated by Syspons. In addition, 25 evaluation reports from 2023 were rated by KEK-CDC, and 60 internal completion notes were rated by SECO-WE programme managers, all using the same framework.

For the 2024 reports, each of the 23 SECO-WE evaluation reports was reviewed by a consultant from Syspons. The main text of the report as well as any available annexes, including in separate documents, were considered during the review. The consultant rated the evaluation report along the six OECD-DAC criteria with the respective sub-criteria. Each sub-criterion was rated on a 4-point scale, based on the assessment of the project according to the evaluation report. Note that to ensure consistency and comparability of the ratings, the rating of the sub-criteria was based on the analysis presented in the report and may differ from the evaluator's assessment of the criteria (provided with the evaluation report). Further, in case no sufficient information was available, the respective sub-criterion was marked "not assessed". In addition to the rating, a brief justification of 1-2 sentences was provided by the consultant referring to 1-3 passages in the evaluation report exemplifying the reasoning behind the rating. Quality assurance followed the same four-eyes system as described above for the REQA assessment.

6 Detailed methodological approach: Thematic Content Analysis and Case Studies

Thematic Content Analysis

At the outset, a coding system was developed based on the central evaluation questions. The system comprised two main strands: one focused on the assessment of Theory of Change and logframe quality, and the other on Business Line-specific analyses, including contextual factors, evidence for causal claims, a thematic analysis of project elements in relation to Business Line Theories of Change, and the alignment between project-level and Business Line-level Theories of Change. The coding structure was accompanied by detailed instructions to ensure consistent application across documents and analysts.

Following this, the evaluation reports were imported into MAXQDA and prepared for analysis. Each document was enriched with background information in a structured Excel file, which facilitated later filtering and intersections. The background variables included project title, year of evaluation, type of evaluation, affiliation to one or more Business Lines, and the assigned REQA and DAC scores. These variables enabled the creation of document sets and supported disaggregated analysis across thematic or methodological dimensions.

Coding was carried out in two iterations. In the first iteration, documents were analysed deductively using the predefined code system. Relevant text passages were assigned to codes according to the coding glossary. In a second, inductive iteration, the analysis was deepened by clustering coded segments into subcodes and refining the coding structure where necessary. Memos were used throughout to document insights on how the subcodes related to the evaluation questions and which specific content and connections they represented.

Finally, the coded data was analysed both qualitatively and quantitatively. Frequency counts helped assess how widely certain themes were covered across reports, while the content of the coded segments was analysed in depth to identify patterns, contrasts, and illustrative examples. In addition, selected codes and thematic aspects were intersected with REQA and DAC scores to explore how evaluation content related to methodological quality or evaluative focus. This combination of deductive structure, inductive refinement, and quantitative validation allowed for a nuanced and systematic synthesis of evaluation content.

Case Studies

The following table shows the ratings (REQA scores and DAC effectiveness rating) for the selected **cases**, which formed the basis for the selection.

Cases	REQA Scores	OECD-DAC Effectiveness
WEHU 226 Evaluation of the Colombia Más Competitiva Programme - Final Report	Average: 2.8 Adequacy: 4 Transparency: 3 Triangulation 3 Appropriateness: 3 Analysis of contribution: 2 Recommendations: 1	Satisfactory

WEMU 121 Revenue Administration and Public Financial Management Reform in Southeast Europe (SEE II) – Mid-term Evaluation	Average: 2.8 Adequacy: 4 Transparency: 2 Triangulation: 2 Appropriateness: 3 Analysis of contribution : NA Recommendations: 3	Unsatisfactory
WEIF 136 Final Evaluation Report – EBRD-SECO ITCP Early-Stage Assessment	Average: 2.25 Adequacy: 3 Transparency: 2 Triangulation: 2 Appropriateness: 2 Analysis of contribution: 2 Recommendations: 2	Highly satisfactory
WEHU 227 External Evaluation – Swiss Trade Policy and Export Promotion Project Vietnam	Average: 1.25 Adequacy: 2 Transparency: 0 Triangulation: 1 Appropriateness: 1 Analysis of contribution: NA Recommendations: 1	Highly Unsatisfactory
WEMU 112 Independent Evaluation of the Extractive Industries Transparency Initiative, Phase IV	Average: 3.0 Adequacy: 4 Transparency: 2 Triangulation 3 Appropriateness: 3 Analysis of contribution: 2 Recommendations: 2	Satisfactory
WEQA Independent evaluation of SECO-WE's climate approach	Average: 2.6 Adequacy: 4 Transparency: 2 Triangulation: 2 Appropriateness: 3 Analysis of contribution: NA Recommendations: 2	Not part of the OECD-DAC rating

7 Detailed findings: Changes in DAC ratings over time

While the projects' performance is found to be mostly satisfactory, previous performance reports generally found higher shares of satisfactory projects. In the years from 2015-2022 less than 5% of projects were assessed unsatisfactory in terms of relevance. Similarly, only 6% of projects were assessed to have unsatisfactory coherence in 2019-2022. For effectiveness and efficiency, the shares of unsatisfactory projects were around 15% in 2015-2022 - more than 10 percentage points less compared to 2023-2024. The projects' impact was however rated lower in 2019-2022 with also 12% with an unsatisfactory rating but some of which highly unsatisfactory. Sustainability in 2015-2018 was rated similar to 2023-2024 and slightly better between 2019 and 2022.[35]

These changes should however not be interpreted as a decrease in performance of SECO-WE's portfolio of projects as statistical uncertainty, selection bias and changes in assessment processes are likely contributors to observed changes. There are several potential reasons for why ratings changed over the years: i) It could reflect a real change in project performance across the portfolio, including e.g. from challenges related to COVID-19 or increased instability in priority countries, ii) it could be caused by the selection of evaluated projects which changes every year, and/or iii) it could be explained by changes in the assessment methodology[36] and the consultants conducting the assessments[37]. Assuming the selection of projects and availability of assessments is random (which is not the case), the margin of error around the share of projects with satisfactory ratings is almost always larger than the observed difference to previous periods. For example, the share of satisfactory projects with regard to effectiveness is 73% with a margin of error of 12 percentage points. Hence, the real share could be somewhere between 61% and 85% (+/- 12%), thus covering the share of 83% of satisfactory projects in 2015-22.[38] So while some differences appear large, they would mostly not be considered statistically significant at the 5% confidence level. Therefore, the selection of evaluated projects would be a sufficient explanation, however, the authors believe another relevant driver of change was that both the assessment methodology and the people conducting the assessments changed over the years as well as the evaluators producing the evaluation reports which the assessments are based on.

[35] SECO-WE (2023). 2021-2022 Performance Report. Meta-analysis of evaluation results from SECO's economic cooperation and development activities. Source: SECO-WE (2023). *2021-2022 Performance Report. Meta-analysis of evaluation results from SECO's economic cooperation and development activities*.

[36] The assessment methodology changed in 2021 substantially (for example, sub-criteria added) and was adapted slightly again in 2023.

[37] Until 2023, the assessments were conducted by KEK-CDC and since 2024 by Syspons. It must be noted that the consultants conducting assessments for KEK-CDC changed over time as well.

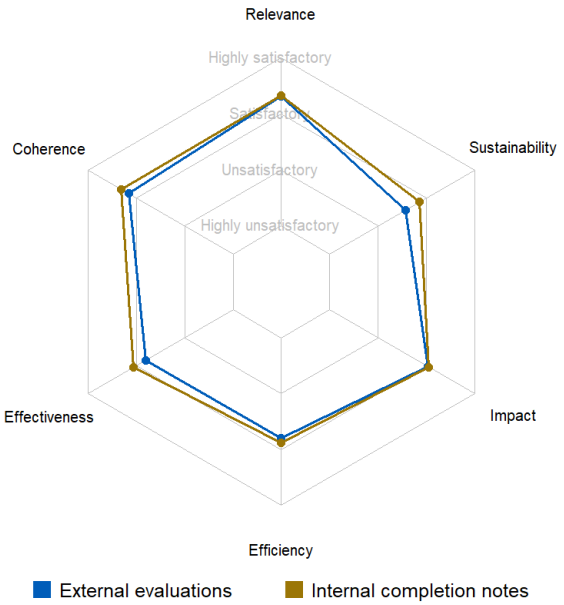
[38] The margin of error is calculated with a finite sample adjustment as $1.96 \cdot \sqrt{\frac{p \cdot (1-p)}{n}} \cdot \frac{N-n}{N-1}$, where n is the number of evaluated projects for the criterion, N the total number of projects, p the share of satisfactory and highly satisfactory projects, and 1.96 is the 97.5% quantile of the Normal distribution, corresponding to a 95% confidence interval.

8 Detailed findings: Internal compared to external assessments

On average, there is little difference between internal and external assessments of projects across OECD-DAC criteria. In addition to mostly external evaluations, some projects are also assessed internally in project completion notes along the same dimensions of the OECD-DAC criteria. Figure 21 displays the average scores across evaluations and completion notes respectively for each OECD-DAC criterion. Note that while there is an evaluation report for 47 projects and a completion note for 60 projects, both is only available for 12 projects. Comparing the two samples of project evaluations, there are only minor differences in ratings between internal and external assessments for the criteria Relevance, Coherence, Impact and Efficiency. Only for Effectiveness and Sustainability the criteria differ[39] and are on average rated slightly lower in the external assessments. Similarly, internal scores also tended to be better compared to external ratings for Coherence, Effectiveness and Sustainability in the years from 2015-2022.[40]

Inspecting the ratings of the 12 projects for which both an internal and external assessments exist, similarly indicates no strong differences. For relevance, coherence, and efficiency, assessments are the same in half the cases and otherwise, internal ratings are similarly likely to be higher or lower than based on the external evaluations. For effectiveness, the rating is the same in half the cases, but when differences occur, external assessments tend to be more critical. Impact assessments generally display a broader divergence, with internal ratings tending to be more negative. In contrast to the overall average score, among the 12 projects with both assessments sustainability tends to be rated worse in internal assessments, especially when compared to mid-term evaluations.

Figure 21 | Internal vs. external assessments



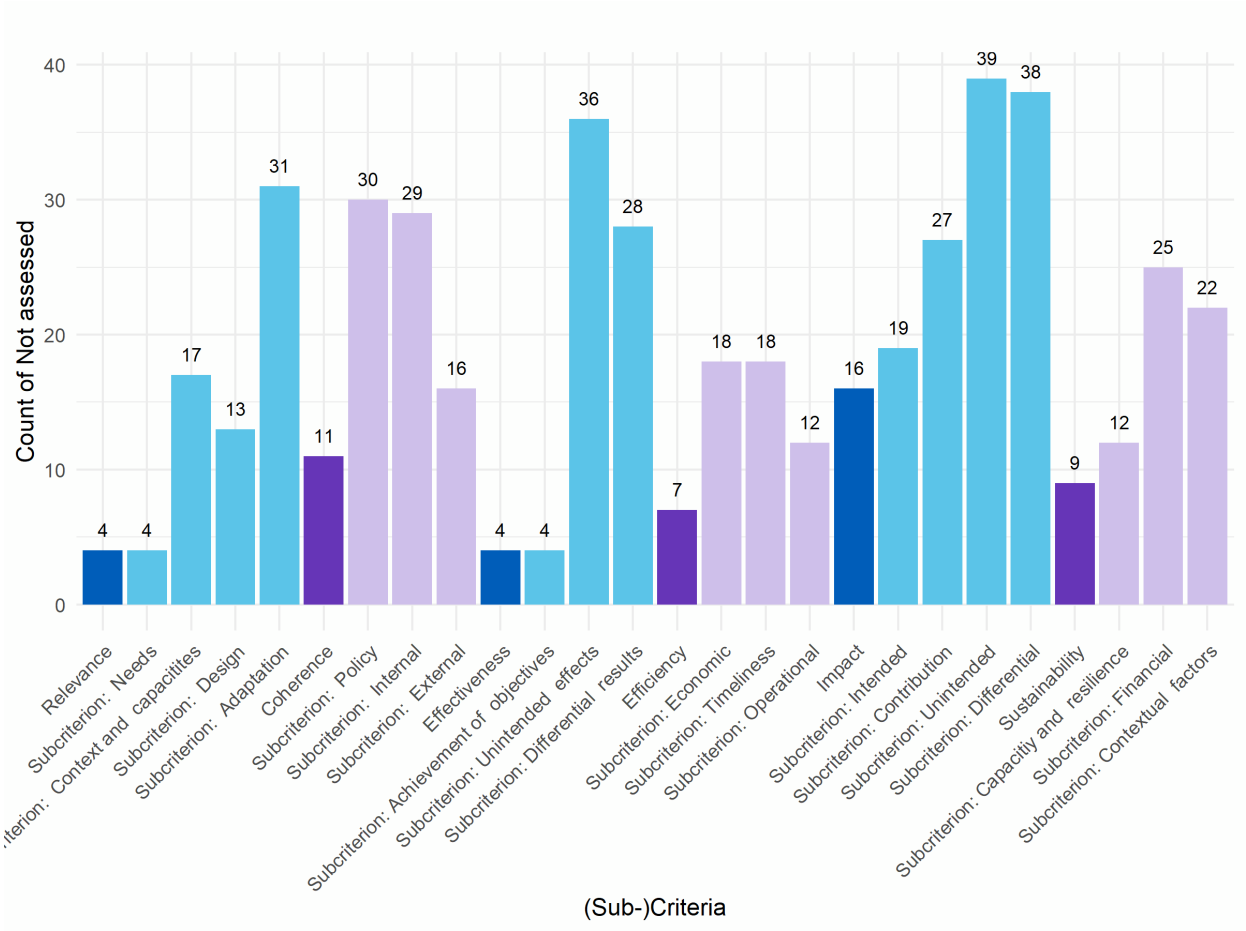
[39] Differences are statistically significant based on a 95% confidence interval of the average difference in ratings between the two samples and a Mann-Whitney U test on the equality of distributions from which the samples originate. However, since the sample of evaluated projects is not random, the significance test needs to be interpreted with caution. Any difference between internal and external evaluations can be due to different projects or differences in assessments

[40] Based numbers reported in: SECO-WE (2023). *2021-2022 Performance Report. Meta-analysis of evaluation results from SECO's economic cooperation and development activities;* Differences are statistically significant based on own calculations of average differences and Mann-Whitney U tests for the equality of distributions of the samples of internal and external assessments. However, since the sample of evaluated projects is not random, the significance test needs to be interpreted with caution.

9 Detailed findings: Availability of information for the assessment along the OECD-DAC criteria

As noted also under limitations (see Chapter 3), the availability and depth of assessment across evaluation criteria vary significantly across the reviewed reports, with some areas receiving more consistent attention than others. While SECO-WE deliberately promotes focused evaluations that do not necessarily cover all DAC criteria, this analysis helps to identify whether certain criteria are being systematically overlooked. **Relevance is generally well covered, though not without gaps.** While the sub-criterion responsiveness to needs, policies and priorities is assessed for 92% of evaluated projects, information related to the alignment of projects with contextual factors and the capacities of target groups is only provided in roughly two-thirds of evaluations. Design considerations are addressed more consistently in about three-quarters of reports. The dimension of adaptability, however, is less frequently assessed in only 35% of projects. **Coherence, particularly internal coherence within the Swiss international cooperation (IC) system, is also less comprehensively addressed.** Less than half of all evaluations include an assessment of internal coherence, with this aspect especially absent in evaluations commissioned by partner organisations. While the sub-criterion policy alignment is addressed in about two-thirds of SECO-commissioned evaluations, it is missing in roughly two out of three evaluations not commissioned by SECO. Internal compatibility is missing in about one-third of these evaluations but available in almost 80% of evaluations commissioned by SECO-WE, underscoring a significant gap in the evaluative coverage of Swiss IC coherence-related dimensions especially in evaluations not commissioned by SECO-WE. Only external compatibility (i.e. compatibility with other donors' or partners' own interventions) is addressed more frequently in 2 out of 3 reports. **Effectiveness assessments are present in most evaluation reports, but many sub-criteria are addressed rather rarely.** Unintended effects are only discussed in one out of every four evaluations, indicating a substantial blind spot. Differential results are assessed in fewer than half of the reports. Gender and inclusion-related analyses appear more frequently, yet they often focus on procedural aspects and highlight the lack of disaggregated results data, limiting their usefulness for assessing actual outcomes. **Efficiency is assessed in most evaluations but based on different sub-criteria.** For each of the efficiency-related sub-criteria, about two-thirds of evaluations provide sufficient information. While individual sub-criteria may not always be covered comprehensively, most reports include at least some relevant content, enabling an overall efficiency assessment in approximately 85% of cases. **Impact is the least assessed criteria and remains one of the most challenging criteria to assess due to limitations in the evidence base.** Around two-thirds of evaluations, for all mid-, end-term and ex-post, include some discussion of impact; however, these are typically restricted to assessing the plausible contribution to or more often the probable achievement of intended impacts based on indicators. Unintended and differential impacts are rarely explored. Fewer than half of the reports address issues of contribution or attribution, and among those that do, rigorous methodological approaches are seldom applied. Moreover, impact discussions are often brief and lack analytical depth, though many reports commendably acknowledge these limitations transparently. **Sustainability is covered unevenly across reports, with notable information gaps.** Capacity-related aspects of sustainability are addressed in about 75% of evaluations, making this the most consistently assessed sub-dimension. In contrast, financial and contextual sustainability are less frequently considered, appearing in only 48% and 54% of evaluations, respectively. When these aspects are addressed, they are often blended into broader discussions, which can obscure findings specific to one sub-criteria.

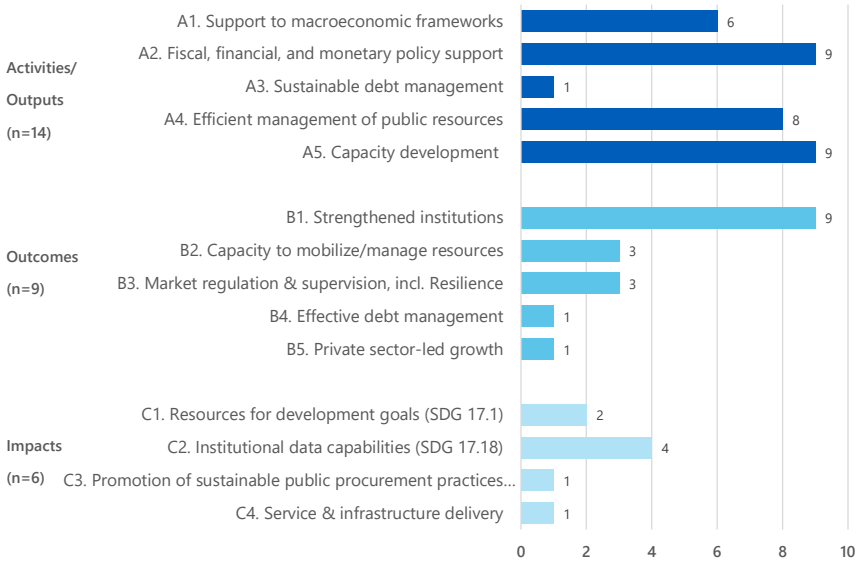
Figure 22 | Number of reports which did not assess different OECD-DAC sub-criteria



10 Detailed findings: Business Line thematic analysis

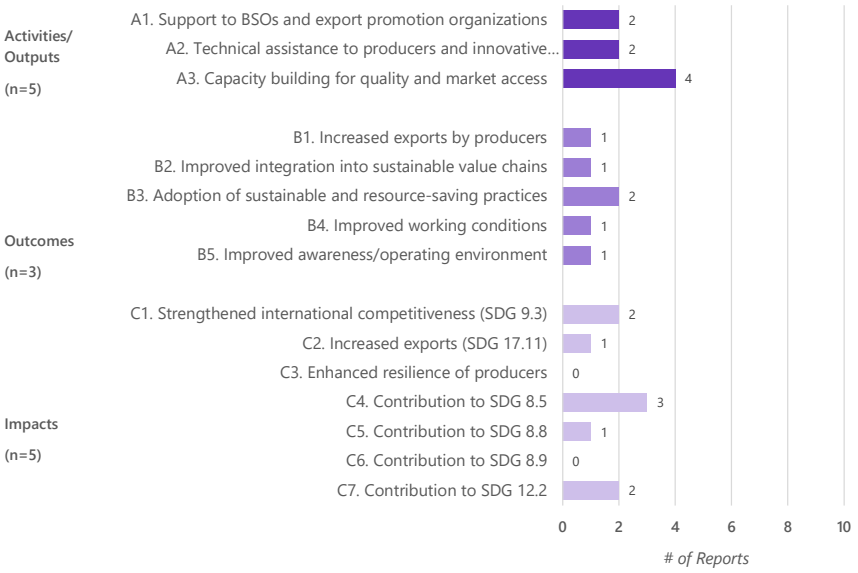
The content analysis suggests that Business Line-specific themes are well represented at the activity and output level, but less consistently traced through to outcomes and impacts. Across all three Business Lines, evaluations often describe what projects did but provide less systematic information on what these interventions achieved or how they contributed to broader change. The depth of analysis decreases noticeably along the results chain. Activities and immediate outputs are far more visible than longer-term outcomes or impact-level contributions. This indicates a general reporting pattern in evaluations that prioritises operational delivery over strategic results analysis.

Figure 23 | Thematic Analysis – BL 1.1 Growth-promoting policies (n=14)



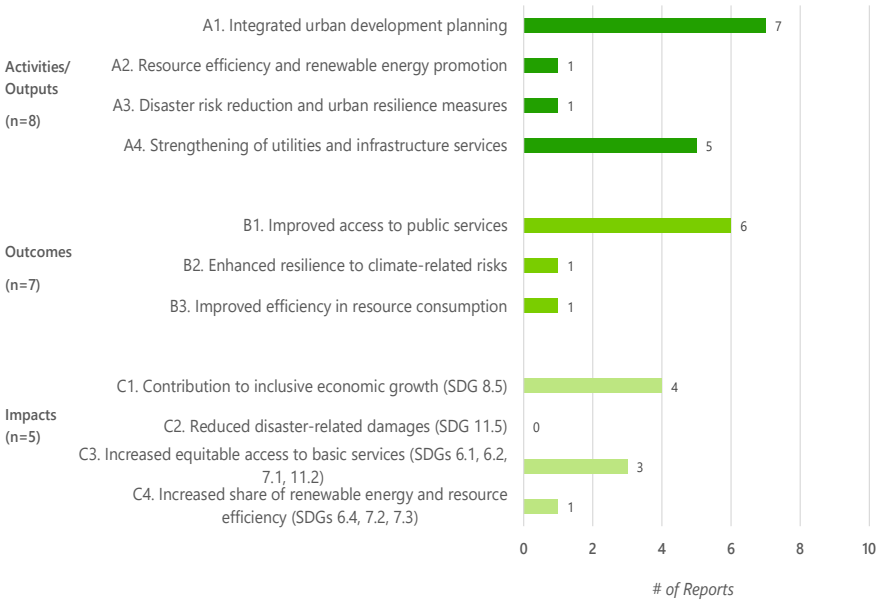
For Business Line 1.1 Growth-promoting Policy, relevant information was found across all three results levels, but with decreasing frequency. All 14 reports included content on activities and outputs, with recurring references to fiscal, financial, and monetary policy support (9 reports), capacity development (9 reports), macroeconomic frameworks (6 reports), and efficient public resource management (8 reports). In contrast, sustainable debt management was only mentioned once. At the outcome level, nine reports contained relevant findings, most of which related to strengthened institutions. Other outcome dimensions - such as improved market regulation or resource mobilisation - were noted only sporadically. Impact-level evidence was found in six reports, with the most common reference being institutional data capabilities (linked to SDG 17.18). Broader themes like improved service delivery or systemic change were rarely documented. This suggests a strong focus on institutional measures, while outcome diversity and impact linkages remain limited.

Figure 24 | Thematic Analysis - BL 2.2 Integration in Value Chains (n=5)



In Business Line 2.2 Integration into Value Chains, thematic elements were identified in all five reports reviewed, but the level of detail and consistency varied considerably across results levels. Activities and outputs were the most commonly reported, particularly capacity-building for market access and quality compliance (4 reports). Other measures, such as export promotion and support to business service organisations, were less frequently mentioned. Outcome-level findings were present in only three reports and mostly concerned the adoption of sustainable production practices. Results such as improved working conditions or increased exports appeared only once each. Impacts were referenced in five reports, often in relation to international competitiveness or decent work (SDG 8.5). However, no report discussed producer resilience - a key theme of this Business Line - suggesting a gap in the thematic coverage. Overall, the analysis indicates that Business Line 2.2 evaluations focus primarily on describing activities, while the tracing of longer-term effects remains incomplete.

Figure 25 | Thematic Analysis - BL 3.1 Urban Development (n=11)



For Business Line 3.1 Urban Development, relevant aspects were identified in most of the eleven reports, but the intensity of coverage declined from activity to impact level. At the activity/output level, eight reports included relevant content, especially on integrated urban planning (7 reports) and infrastructure strengthening (5 reports). More specific measures, such as energy efficiency, renewable energy promotion, or disaster risk reduction, were mentioned only once each. Outcome-level evidence was found in seven reports, most commonly in relation to improved access to public services (6 reports). Other outcomes, such as resilience to climate-related risks or resource efficiency, were only sporadically addressed. Impact-level contributions were identified in five reports, with inclusive growth and equitable access to services being the most frequent themes. However, key impact areas like reduced disaster-related damages or environmental sustainability received little to no attention. This suggests that while the planning and service delivery dimensions of Business Line 3.1 are well covered, other thematic priorities, particularly in relation to resilience and climate, are not consistently reflected in the evaluations.

11 Additional Figures and Tables

Figure 26 | Overview of analysed evaluation reports per year and section

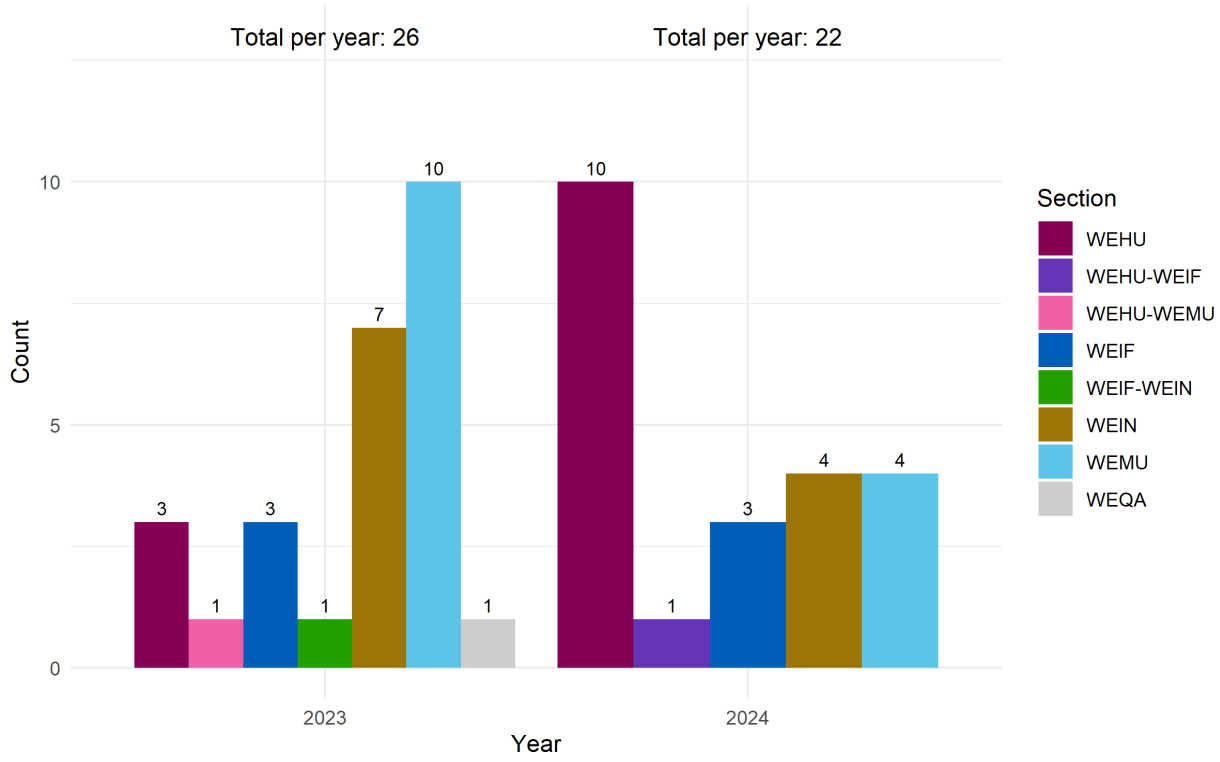


Figure 27 | Internal correlation REQA scores

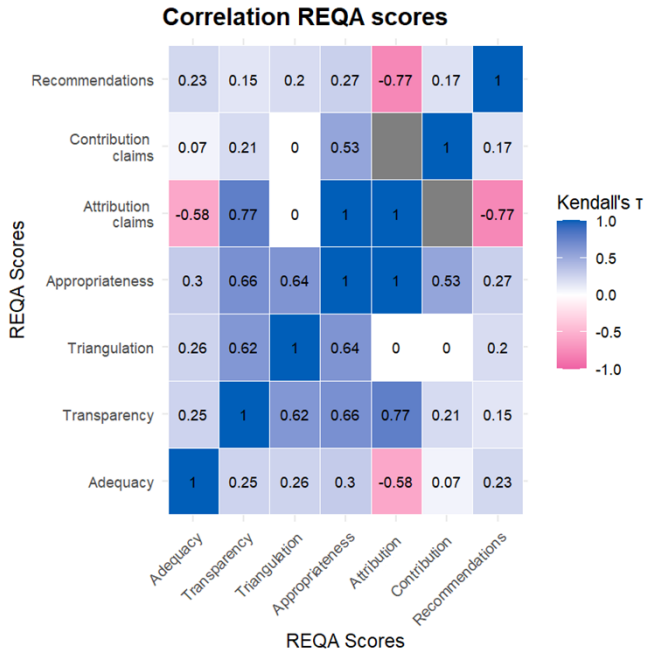


Figure 28 | Average REQA rating per year

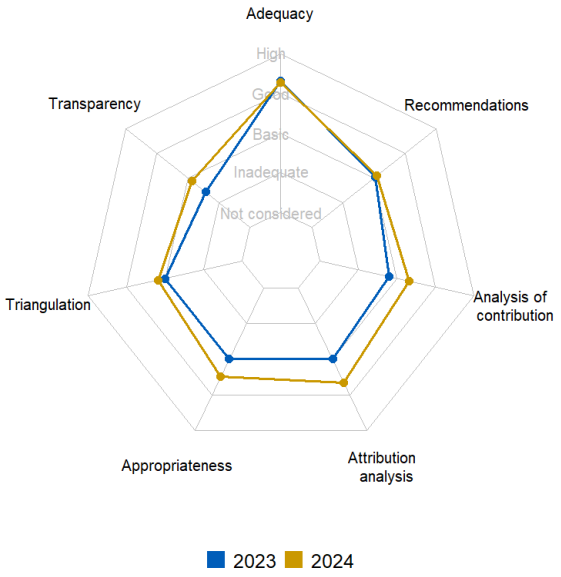


Figure 29 | Average REQA rating by section (note does not include evaluations assigned to more than one section)

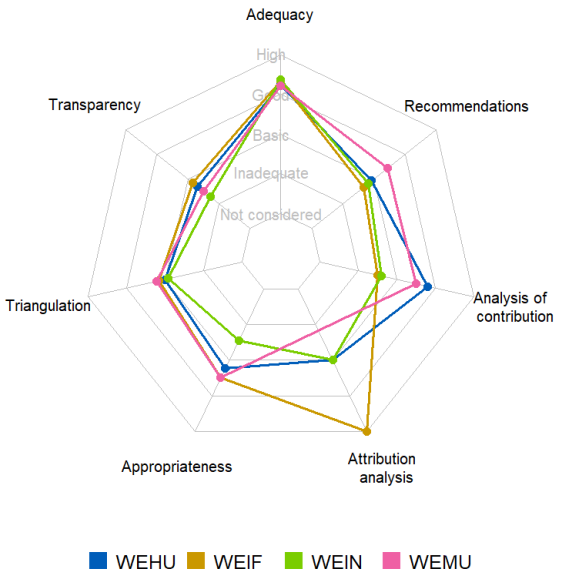


Figure 30 | Overview OECD-DAC ratings

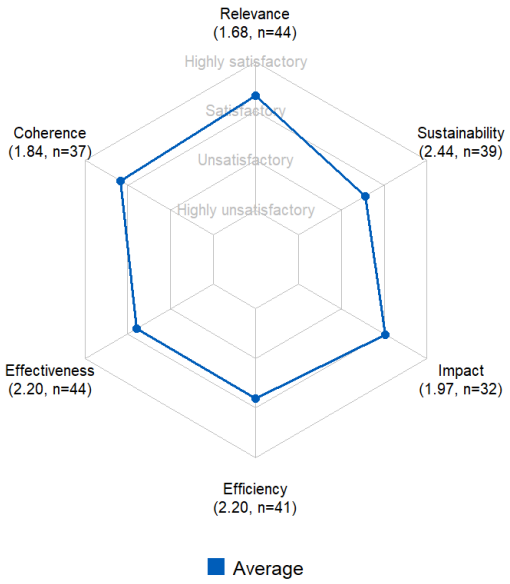


Figure 31 | Average DAC rating by year

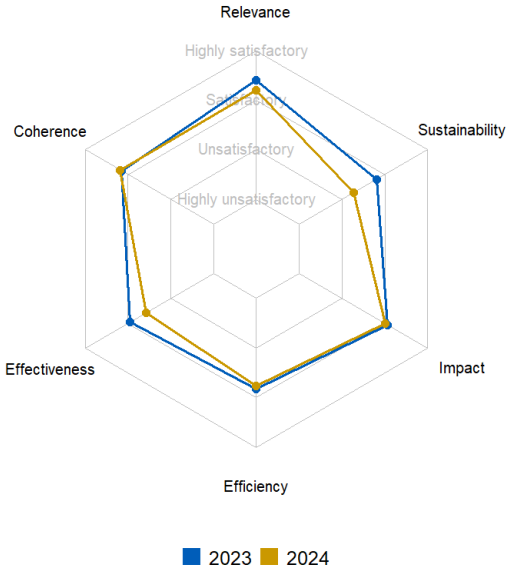


Figure 32 | Average DAC rating by section

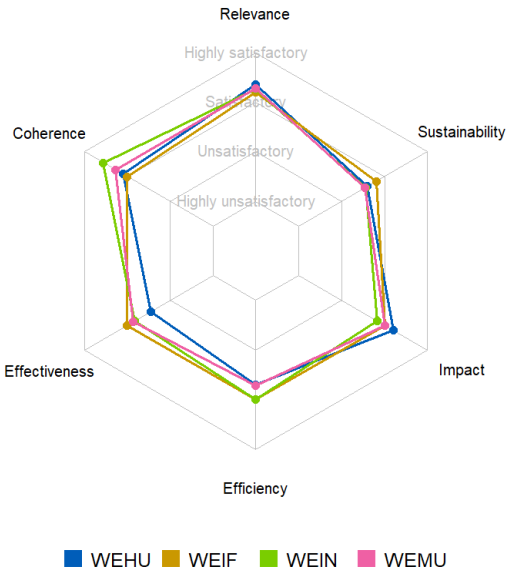
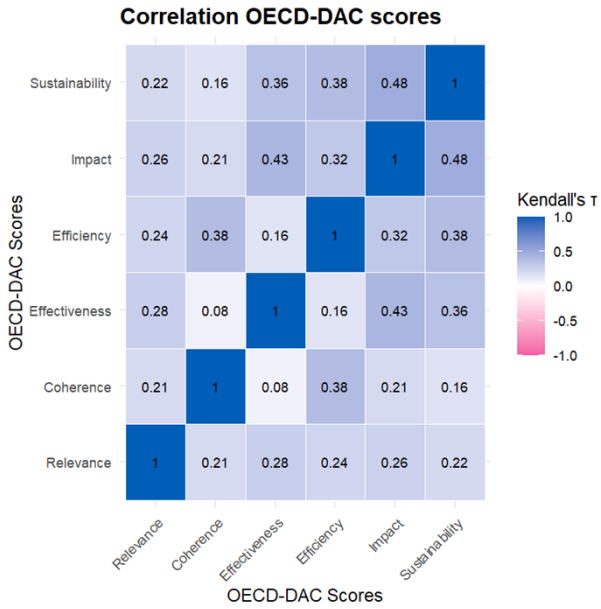


Figure 33 | Correlation OECD-DAC scores



Intersection of DAC scores with coded assessments of Theory of Change quality

Table 4: Frequency Table: Relevance – subcriterion: quality of design x Theory of Change Quality

ToC Quality Assessment	Not assessed	Highly satis-factory	Satisfactory	Unsatisfactory	Highly unsatis-factory
no judgement	2	1	5	0	0
as evaluators designed ToC themselves, no judgement	2	3	2	0	0
judged as clear and useful	1	0	2	0	0
judged as vague or incom-plete	3	1	2	5	0
ToC not used or deemed not useful	0	0	0	0	0

Table 5: Frequency Table: Effectiveness (overall) x Theory of Change Quality

ToC Quality Assessment	Not assessed	Highly satis-factory	Satisfactory	Unsatisfactory	Highly unsatis-factory
no judgement	0	0	4	4	0
as evaluators designed ToC themselves, no judgement	0	2	4	1	0
judged as clear and useful	0	0	3	0	0
judged as vague or incom-plete	1	0	7	2	1
ToC not used or deemed not useful	0	0	0	0	0

Table 6: Frequency Table: Effectiveness – subcriterion: attainment of objectives x Theory of Change Quality

ToC Quality Assessment	Not assessed	Highly satis-factory	Satisfactory	Unsatisfactory	Highly unsatis-factory
no judgement	0	1	5	2	0
as evaluators designed ToC themselves, no judgement	0	3	3	1	0
judged as clear and useful	0	0	3	0	0
judged as vague or incom-plete	1	1	6	2	1
ToC not used or deemed not useful	0	0	0	0	0

Table 7: Frequency Table: Impact (overall) x Theory of Change Quality

ToC Quality Assessment	Not assessed	Highly satis-factory	Satisfactory	Unsatisfactory	Highly unsatis-factory
no judgement	2	2	3	1	0
as evaluators designed ToC themselves, no judgement	0	1	5	1	0
judged as clear and useful	1	0	2	0	0
judged as vague or incom-plete	6	0	5	0	0
ToC not used or deemed not useful	0	0	0	0	0

Table 8: Frequency Table: Impact – subcriterion: intended impacts x Theory of Change Quality

ToC Quality Assessment	Not assessed	Highly satis-factory	Satisfactory	Unsatisfactory	Highly unsatis-factory
no judgement	3	0	5	0	0
as evaluators designed ToC themselves, no judgement	0	3	3	1	0
judged as clear and useful	1	0	2	0	0
judged as vague or incom-plete	7	1	3	0	0
ToC not used or deemed not useful	0	0	0	0	0

Intersection of DAC scores with coded assessments of Logframe quality

Table 9: Frequency Table: Effectiveness (overall) x Logframe Quality

Logframe Quality Assessment	not assessed	highly satisfactory	satisfactory	unsatisfactory	highly unsatisfactory
judged as clear and adequate	0	0	9	4	0
judged as deficient or inadequate	0	3	10	4	1

Table 10: Frequency Table: Effectiveness – subcriterion: attainment of objectives x Logframe Quality

Logframe Quality Assessment	Not assessed	Highly satisfactory	Satisfactory	Unsatisfactory	Highly unsatisfactory
judged as clear and adequate	0	3	7	3	0
judged as deficient or inadequate	0	4	9	4	1

Table 11: Frequency Table: Impact (overall) x Logframe Quality

Logframe Quality Assessment	Not assessed	Highly satisfactory	Satisfactory	Unsatisfactory	Highly unsatisfactory
judged as clear and adequate	1	3	6	3	0
judged as deficient or inadequate	11	0	6	1	0

Table 12: Frequency Table: Impact – subcriterion: intended impacts x Logframe Quality

Logframe Quality Assessment	Not assessed	Highly satisfactory	Satisfactory	Unsatisfactory	Highly unsatisfactory
judged as clear and adequate	1	3	6	3	0
judged as deficient or inadequate	12	2	3	1	0

Table 13: Frequency Table: Effectiveness (overall) x Reflections on logframe–effectiveness relationship

Reflections on logframe–effectiveness relationship	Not assessed	Highly satisfactory	Satisfactory	Unsatisfactory	Highly unsatisfactory
Positive Link	0	0	1	1	0
Negative Link	0	1	3	0	0

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs,
Education and Research EAER
State Secretariat for Economic Affairs SECO

SYS
PONS

CONTACT

State Secretariat for Economic Affairs (SECO)
Economic Cooperation and Development Division
Evaluation Unit (WEQA)

we.evaluation@seco.admin.ch

Syspons GmbH

info@syspons.com