Review of the quality and use of DFAT evaluations: 2024-2025

Submitted to the
Department of Foreign Affairs and Trade
by Bluebird Consultants

19 December 2025

Enquiries:
Jessica Kenway, Director
M: 0425 878 368
E: jess@bluebirdconsultants.com.au
www.bluebirdconsultants.com.au

bluebird consultants

Contents

Acknowledgements

The review was conducted by a team of 3, led by Ms Clare Strahan and with core contributions from Mr Joseph Thompson and Ms Jenny Geppert. Ms Rachel Ingwersen from the Development Evaluation and Assurance Section conceptualised the review and ensured it met DFAT standards.

This review was funded by the Australian Department of Foreign Affairs and Trade but does not represent the views of government. Any omissions or inaccuracies remain the responsibility of the team.

List of Acronyms

Acronym	Meaning
AAAEP-P	Australia Awards and Alumni Engagement Program – Philippines
AUD	Australian Dollar
DFAT	Department of Foreign Affairs and Trade
DMEL	Design, Monitoring, Evaluation and Learning
EIS	Evaluation Improvement Strategy
EOPO	End of Program Outcome
EVS	Development Evaluation and Assurance Section
FWCC	Fiji Gender Country Program
GEDSI	Gender Equality, Disability and Social Inclusion
KASfEP	Kiribati Australia Skills for Employment Program
KEQ	Key Evaluation Question
MEL	Monitoring, Evaluation and Learning
MTR	Mid-Term Review
PADC	Performance of Australian Development Cooperation
P&Q	Performance and Quality
PRD	Development Effectiveness and Enabling Division
QA	Quality assurance
QCA	Qualitative Comparative Analysis
SAWASI	South Asia Water Security Initiative
TEIP	Targeted Evaluation Improvement Program
ToR	Terms of Reference
WFF MTR	Women's Fund Fiji Mid-Term Review

Executive summary

Every year, the Department of Foreign Affairs and Trade (DFAT) reviews the quality of its development evaluations, based on its 2023 International Development Policy commitment to performance management. **In 2025, the Annual Review focuses on evaluation plans. This reflects a key finding from the 2024 review that staff reported least confidence in reviewing evaluation plans and assessing their quality against the DFAT Design and Monitoring, Evaluation and Learning (DMEL) Standards.** The Development Evaluation and Assurance Section (EVS) wanted to investigate the link between the quality of plans and reports and test the hypothesis that strengthening the quality of plans could strengthen the overall quality of evaluations.

The review set out to answer the following questions:

1. To what extent are DFAT evaluation plans good quality (meeting DMEL Standard 9)
    a. What are the common strengths and weaknesses of the evaluation plans assessed?
    b. What is the correlation between the quality of the evaluation plans and the quality of the final reports?
2. What are the characteristics of good and poor-quality evaluation plans?
3. What practical actions are recommended to improve and strengthen the evaluation plans?

To answer these questions, the 2025 Annual Review of the quality and use of evaluation plans assessed the quality of 20 DFAT-led evaluation plans completed between January 2024 and June 2025. A sample of related evaluation reports were also assessed. The Review explored common strengths and weaknesses and the relationship between plan quality and report quality, using a sample of nine low-, medium- and high-quality evaluation plans.

The review found only eight of the 20 evaluation plans (40%) were of adequate quality or above (Q1) when reviewed against the DFAT Design Monitoring Evaluation and Learning (DMEL) Standard 9. Common weaknesses centred on collaborative approaches, processes and approaches for generating evidence including sampling; analytical approaches; limited use of monitoring and secondary data; activity planning and scheduling; and gender equality, disability and social inclusion (GEDSI) considerations. Common strengths included clear purpose, evaluation questions, and defined roles and responsibilities. These areas still had room for improvement, such as too many sub-questions and limited prioritisation among evaluation questions (Q1a).

At a macro level, there was a limited high-level correlation between the quality of plans and reports given high-quality reports were linked to low-, medium- and high-quality plans (Q1b). But a more granular analysis showed **higher quality plans – specifically those tailored well to context and demonstrating stronger analytical approaches – were linked to reports with stronger evidence.** While the small sample size means the findings are indicative rather than conclusive, they still provide valuable insights to guide actions and recommendations supporting staff to review evaluation plans and better align them with DFAT Standards.

The review found good-quality plans were clear on the purpose and intended use of the evaluation (Q2). The evaluation plan focused on a limited set of evaluation questions, tied closely to the program's objectives with attention paid to GEDSI. Good-quality plans provided

sufficient methodological detail, outlining sampling strategies, processes for data analysis and appended data collection tools, as well as making appropriate use of monitoring and secondary data. These plans were based on realistic schedules with clearly defined roles and responsibilities, including QA, and acknowledged any limitations in evaluation design alongside responses to these limitations.

Poor-quality plans generally provided little detail on how findings would be used. These plans tended to lack key design elements and paid limited attention to ethics or flexibility. Poor-quality plans insufficiently explained sampling strategies and made little use of monitoring or secondary data. Data collection tools were often underdeveloped or not included, and analytical approaches tended to be weak. There was limited use of triangulation and no clear process to judge the strength of evidence. This made it hard for the reader to understand how findings and conclusions would be reached.

This review has made three recommendations (Q3).

1. EVS to consider focusing on weaker elements in plans for additional support and training. These areas (listed below) could be elaborated on through existing tools and mechanisms. It is recommended EVS discuss the topics and create an Action Plan for implementation.

- Enabling collaborative approaches to evaluations
- Developing Key evaluation questions (KEQs) and sub-questions
- Developing appropriate sampling strategies
- Demonstrating strength of evidence within evaluations
- Aligning limitations, findings and conclusions
- Defining approaches to data analysis and judgement-making
- Embedding GEDSI considerations throughout the evaluation cycle

2. EVS should clarify and communicate expectations for DFAT staff and contractor use of DMEL Standards. This could be part of the Action Plan from Recommendation 1.

- Outline strategies to clearly define and communicate DFAT's expectations for contractor use of the DMEL Standards. This could include industry briefings
- Support DFAT staff to apply the Standards when reviewing Terms of Reference (ToR), plans and reports
- Communicate how DFAT staff can draw on existing support to put these improvements in place
- Work with DFAT procurement to further embed the Standards.

3. EVS to consider revising the DMEL Standards to improve clarity and to emphasise focus areas. A number of Standards are noted for revision under Terms of Reference (Standard 8), Evaluation Plan (Standard 9), and Evaluation Report (Standard 10), with details included in the report.

1. Introduction

The Development Evaluation and Assurance Section (EVS) of the Department of Foreign Affairs and Trade (DFAT) commissioned Bluebird Consultants (Bluebird) to review a sample of its 2024-25 development evaluation plans and associated reports, focusing on quality. This review followed five previous reviews of evaluation quality in 2012, 2014, 2017, 2023 and 2024.

2. Background

The Australian Government committed to strengthening monitoring, evaluation and learning (MEL) approaches in its 2023 International Development Policy. One indicator of the Performance and Delivery Framework, which underpins this policy, is as follows: "*Our development cooperation is informed by monitoring, evaluation and learning"*. Specifically, DFAT has committed to conducting an annual review of the quality and use of evaluations and to publicly report on the findings as a measure of this indicator. This review has provided data contributing to this commitment.

DFAT completes on average around 40 development evaluations each year, in line with the Development Evaluation Policy (also referred to as 'the policy' in this report), which was introduced in 2016 and updated in 2023. Each prior quality review has had a slightly different focus, as follows:

- 2012: Reviewed evaluation practices and the quality of independent evaluations.
- 2014: Assessed evaluation quality and the factors influencing it and provided improvement recommendations.
- 2017: Examined evaluation quality, utility and the impact of the 2016 Aid Evaluation Policy and identified lessons on policy influence, capability and gender equality.
- 2023: Reviewed evaluation report quality and analysed the use of recommendations.
- 2024: Assessed how effectively findings were used.

A key finding from the 2024 review was that the DFAT staff sampled reported least confidence in reviewing evaluation plans and reports against the DMEL Standards. DFAT's EVS assumed that this was related to several factors, including uncertainty around who was responsible for the quality review of evaluation plans, the varying levels of confidence in technically assuring evaluation plans, and the limited time available to conduct thorough quality assurance processes of plans and reports. Similar insights were also gathered from the DFAT staff who were supported through the Evaluation Helpdesk and Targeted Evaluation Improvement Program (TEIP). Consequently, EVS identified that it would be useful to focus this year's review on the quality of evaluation plans and the extent to which, and how, this may influence the quality of evaluation reports.

3. Purpose

Primarily, this review provided an opportunity to assess how well DFAT is implementing a key element of the Development Evaluation Policy: quality (specifically, in relation to evaluation plans). The review explored common strengths and weaknesses in the evaluation plans,

examined the correlation between the quality of the evaluation plans and the final reports, identified characteristics of strong and weak evaluation plans, and suggested practical actions to strengthen them.

The review was conducted in a way that built EVS capability through close management of the work.

4. Objectives

The review had the following three objectives:

1. Contribute to the measurement of the Tier 3 indicator in the Performance and Delivery Framework for the following International Development Policy indicator: "*Our development cooperation is informed by monitoring, evaluation and learning*". The measure is as follows: "*conduct an annual review of the quality and use of evaluations and publicly report on the findings*".
2. Support future assessments of the progress and impact of the Evaluation Improvement Strategy (EIS), through establishing a baseline against which to assess the support it has given.[1]
3. Track the quality and use of evaluation outputs over time to inform DFAT policy and practice.

5. Scope

The review was limited to DFAT-led evaluations (those commissioned and managed by DFAT), which were completed between 1 July 2024 and 30 June 2025. The review sampled 20 evaluation plans from the 37 DFAT-led program evaluations that were completed during this period. The purposive sampling used the following criteria: geographic coverage, investment value and sector. This enabled diversity and breadth in line with the investment emphasis. Nine evaluation reports were selected for review, including three that were assessed as having low-, medium- and high-quality evaluation plans. A comparative review was undertaken across the evaluation plans and associated reports to explore the relationship between plan quality and report quality.

6. Intended use and intended users

The primary users of this review will be Senior Managers in the Development Effectiveness and Enabling Division (PRD), PRD staff working on design and procurement, and EVS. The findings from this review will be used by these groups to inform their approaches, policies, guidance, training and practices in MEL. The review is also expected to have broader utility for all staff working on the development program.

The DFAT staff involved in commissioning and managing evaluations, the performance and quality focal points, and the monitoring and evaluation advisers will be interested in the review's findings to improve the commissioning, support, management and use of higher quality evaluations. More broadly, the Australian public will also be an audience, as the

[1] The baseline only applies to evaluations that have not yet received support from the EIS.

published review will contribute to understanding the performance of the development program.

The quality reviews of plans (and the associated scores) will also be considered as a baseline against which to consider the EIS's impact, as the review predates it.

The findings of this review will be shared with DFAT staff and other stakeholders through:

- Publication of key findings in the Performance of Australian Development Cooperation (PADC) Report 2024-2025.
- Publication of the full review report, including the executive summary, on the DFAT website.
- Presentation of the review process and findings by EVS staff and Bluebird (with DFAT's permission) at appropriate forums within DFAT and externally as opportunities arise, e.g., with DFAT's Performance & Quality (P&Q) network, at EvalNet meetings, and with companies and consultants who undertake DFAT evaluations.

7. Review questions

The review answered the following questions:

1. To what extent are DFAT evaluation plans good quality (meeting DMEL Standard 9)?
    a. What are the common strengths and weaknesses of the evaluation plans assessed?
    b. What is the correlation between the quality of the evaluation plans and the quality of the final reports?
2. What are the characteristics of good and poor-quality evaluation plans?
3. What practical actions are recommended to improve and strengthen the evaluation plans (considering all the tools available, including the EIS, the Standards, and the Development Evaluation Policy)?

8. Methodology

This review used a mixed-methods design, including quality reviews of both the plans and reports, which were guided by a pre-determined template and both quantitative and qualitative data analytical approaches. The review was conducted over **three stages:** Stage 1 was focused on the quality review of plans, Stage 2 on the quality review of reports, and Stage 3 on analysis and reporting.

A **purposive sample of 20 evaluation plans** was selected from the wider sample of 37 DFAT-led evaluations. The sampling approach looked to ensure diversity and breadth in line with DFAT's investment emphasis, with specific variables of interest, including geographic coverage, sector and value of investment.[2] This is outlined in

[2] DFAT tends to consider investments below AUD 10 million as low-value investments, which relates to the thresholds for increasing quality requirements and levels of approval at the design stage (i.e., peer review, independent appraisal, and Development Program Committee approval).

Table 1.

Table 1: Summary of sampled evaluations for quality review of plans

PADC Categorisation	Sector	Investment Value
Pacific: 8 Southeast Asia: 5 South and Central Asia: 2 Beyond the Indo-Pacific: 1 Global: 1 Sector: 3	Governance: 9 Humanitarian: 1 Agriculture, Trade and other Production Sectors: 3 Education: 1 Health: 3 Economic Infrastructure and Services: 2 Cross-sector: 1	Low (<AUD$10m): 1 Medium (AUD$10–100m): 13 High (>AUD$100m): 6

The **quality review template** (Annex 2) for the plans was developed based on DFAT DMEL Standard 9 and for the reports based on DFAT DMEL Standard 10. The template focused on similar quality criteria across the two sets of standards, which enabled a review of possible correlations between plans and reports. Ratings are included in **Table 2**, below.

Table 2: Ratings

Rating	Satisfactory	Rating	Less than satisfactory
6	**Very high quality:** satisfies criteria in all areas.	**3**	**Less than adequate quality**: on balance, does not satisfy criteria and/or fails in at least one major area.
5	**Good quality:** satisfies criteria in almost all areas.	**2**	**Poor quality:** does not satisfy criteria in several major areas.
4	**Adequate quality**: on balance, satisfies criteria and does not fail in any major area.	**1**	**Very poor quality**: does not satisfy criteria in any major area.

Following the review of the plans, the review team selected a sub-sample of nine plans to conduct a **quality review of associated evaluation reports** (

Table 3). The aim was to select three each from the high, adequate and unsatisfactory quality bands, based on the average rating.

Table 3: Summary of sampled evaluations for the report quality review

Quality category	PADC Categorisation	Sector	Investment Value
Low	Pacific: 1 Sector: 2	Governance: 1 Agriculture, Trade and other Production Sectors: 1 Health: 1	Medium: 2 High: 1
Medium	Pacific: 1 Southeast Asia: 1 South and Central Asia: 1	Governance: 1 Education: 1 Economic Infrastructure and Services: 1	Low: 1 Medium: 2
High	Pacific: 1 Southeast Asia: 1 South and Central Asia: 1	Governance: 1 Agriculture, Trade and other Production Sectors: 1 Education: 1	Medium: 3

The review used the following analytical approaches:

- **Analysis of evaluation plans**
  - **Quantitative:** Assessed average ratings, ranges and distributions across quality criteria; conducted frequency counts of strengths and weaknesses based on up to five identified sub-criteria per plan; and undertook a pivot analysis to explore any correlation between plan quality and geographic coverage, sector or value.
  - **Qualitative:** Thematic analysis by quality criteria to identify common strengths and weaknesses, supported by internal discussions to synthesise cross-plan reflections.
- **Analysis of evaluation reports**
  - **Quantitative**: As with the plans, examined the average ratings, variation across quality criteria and frequency of strengths and weaknesses.
  - **Qualitative:** Thematic analysis of strengths and weaknesses and internal discussion to consider patterns and any links between the plan and report quality.
- **Qualitative Comparative Analysis (QCA)**
  - Explored the alignment between plan and report ratings, shared strengths and weaknesses, and shared factors influencing whether higher quality plans led to higher quality reports using the nine matched plan–report pairs across high-, medium- and low-quality categories.
- **Corroboration and validation**
  - Findings from across analytical components and proposed recommendations were reviewed in a team workshop with the EVS Review Manager and Quality Assurance lead.
  - Key characteristics of high- and low-quality plans were profiled, with case examples developed to illustrate quality components.

The following two key **limitations** applied to the review:

1. The review drew on a small sample of evaluation reports, which limited the extent to which linkages with plans could be explored. As a result, any correlations or key findings were necessarily tentative and are presented with appropriate caveats throughout the report.
2. No interviews were undertaken with Investment Managers or Evaluators, which may have provided deeper insight into the context and factors influencing the plan and report quality. The team reflected on possible influencing factors, but these are presented as hypotheses. A corroboration approach drawing on multiple analytical methods was used to examine quality from different perspectives.

9. Findings

9.1 Quality of the evaluation plans

To what extent are DFAT evaluation plans good quality (DMEL Standard 9)?

Overall, the **plans varied from good to poor quality**. **Only eight plans were assessed as adequate or good quality** (above 4), with only one plan specifically assessed as good quality (above 5). The heat map in **Table 4** presents the average ratings (across all quality criteria) across the 20 plans, presented from the lowest rating (2.1) to the highest rating (5.4).[3] There is **variability in quality both within and across reports**, although overall high-scoring plans tended to score higher across most criteria. **Table** 5 provides a summary of how plan ratings for each quality criteria were distributed.

Table 4: Heat map: Plan ratings by quality criteria, based on DMEL Standard 9

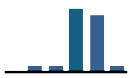
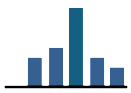
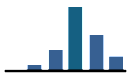
This table uses red to green shading to provide a heat map overview of evaluation plan ratings, where low scores of 1 are dark red and strong scores of 6 are dark green. A version of this table summarising data distribution without colour shading is available in Annex 5, **Table K.**

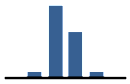
Key quality areas	Plan 1	Plan 2	Plan 3	Plan 4	Plan 5	Plan 6	Plan 7	Plan 8	Plan 9	Plan 10	Plan 11	Plan 12	Plan 13	Plan 14	Plan 15	Plan 16	Plan 17	Plan 18	Plan 19	Plan 20
1) Purpose and use of evaluation	2	5	4	4	4	5.5	4	4.5	4	3.5	5.5	4	4	4	5	5	5	5.5	5	6
2) Evaluation design	2	2.5	2.5	3	4	5	4	4.5	3	4.5	4.5	5	5	4.5	3	4	3	6	4	6
3) Evaluation questions	4	4.5	4	4	3	3.5	4	4	4	5	5	2.5	4	5.5	5	5	6	3.5	4	6
4) Strength of evidence	3	2.5	3	4	3	4	3	3	3	4	4	3	4	4	3	4.5	5	4	5	5

[3] These are a simple average score, not weighted across quality criteria.

Key quality areas	Plan 1	Plan 2	Plan 3	Plan 4	Plan 5	Plan 6	Plan 7	Plan 8	Plan 9	Plan 10	Plan 11	Plan 12	Plan 13	Plan 14	Plan 15	Plan 16	Plan 17	Plan 18	Plan 19	Plan 20
5) Analytical approach	1	3	4	2	3.5	2	3	1.5	2	3	2	4.5	4	3	3	4	2	4.5	5	4
6) Limitations	1	3	3.4	3	4	3.5	3	4.5	6	3	4	5	4	4.5	4	3.5	6	4	5	6
7) Activity planning and scheduling	2	3	2.5	4	3.5	4	4	3.5	4	3.5	2.5	3	4	3	5	5	5	3	6	5
8) Roles and responsibilities	2	4	4	4	4	2.5	5	4.5	4	4	4	4	4	4	5	5	5	6	5	5
Overall plan review score	**2.1**	**3.4**	**3.4**	**3.5**	**3.6**	**3.8**	**3.8**	**3.8**	**3.8**	**3.8**	**3.9**	**3.9**	**4.1**	**4.1**	**4.1**	**4.5**	**4.6**	**4.6**	**4.9**	**5.4**

Table 5: Distribution analysis: Summary of plan ratings by quality criteria, based on DMEL Standard 9

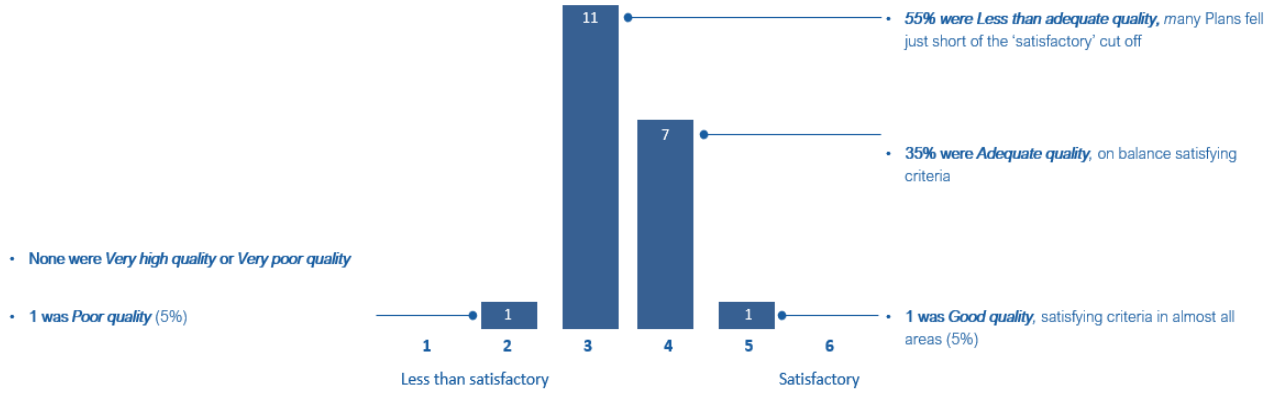
Key quality areas	Min	Max	Mean	Mode	Distribution	Less than satisfactory (0-3)	Satisfactory (4-6)
1) Purpose and use of evaluation	2	6	4.5	4		10%	90%
2) Evaluation design	2	6	4.0	3		35%	65%
3) Evaluation questions	2.5	6	4.3	4		20%	80%
4) Strength of evidence	2.5	5	3.7	3		45%	55%
5) Analytical approach	1	5	3.1	3		65%	35%
6) Limitations	1	6	4.0	4		40%	60%
7) Activity planning and scheduling	2	6	3.8	4		50%	50%

Key quality areas	Min	Max	Mean	Mode	Distribution	Less than satisfactory (0-3)	Satisfactory (4-6)
8) Roles and responsibilities	2	6	4.3	4		10%	90%
Overall plan review score	**2.1**	**5.4**	**3.9**	-		**60%**	**40%**

Sixty percent (12/20) of plans were assessed as below adequate quality (rated 1-3).

Figure 1 shows the most common rating was 3 (less than adequate quality).

Figure 1. Bar chart: Plan ratings across quality criteria, based on DMEL Standard 9



There was no significant variation in plan quality by **geographic area**[4]. In terms of **sector**, there was some variability across categories, indicating that sector was not a strong factor influencing quality. Most sampled plans were in the governance sector (9 out of 20) and seven of these were assessed as unsatisfactory.

In terms of **investment value,** most of the 20 plans reviewed were for medium or higher value investments (13 and 6, respectively). Five out of the six higher value investment plans were rated as unsatisfactory, which is notable given that higher value investments generally have higher evaluation budgets and may be higher-stakes evaluations. Summaries of the pivot analysis are included in Annex 3.

9.2 Common areas of strengths and weaknesses in evaluation plans

Weaknesses of the evaluation plans assessed

The three weakest areas across the evaluation plans, according to the average ratings, were the analytical approach, activity planning and scheduling, and the strength of evidence (

Table 6 heat map and **Table 7** distribution summary). More specifically, the poorest elements of evaluation plans were:

[4] See Table 3 for geographic and sector categories.

- Lack of elaboration on data collection and analysis approaches.
- Missing or low-quality evaluation tools.
- Missing or low-quality processes for assessing the strength of evidence and making judgements to inform conclusions.
- Sampling across data sources not well considered.
- Weak articulation of triangulation and corroboration processes.
- Limited consideration of respondents' needs and rights.
- Little attention given to security of data.
- Micro planning of evaluation activities, specifically time considerations, lacked detail.

Table 6: Heat map: Ratings applied to plans across the three lowest scoring areas

This table uses red to green shading to provide a heat map overview of evaluation plan ratings, where low scores of 1 are shaded dark red and strong scores of 6 are dark green. A version of this table summarising data distribution without colour shading is available in Annex 5, ***Table L****.*

Key quality areas	Plan 1	Plan 2	Plan 3	Plan 4	Plan 5	Plan 6	Plan 7	Plan 8	Plan 9	Plan 10	Plan 11	Plan 12	Plan 13	Plan 14	Plan 15	Plan 16	Plan 17	Plan 18	Plan 19	Plan 20
4) Strength of evidence	3	2.5	3	4	3	4	3	3	3	4	4	3	4	4	3	4.5	5	4	5	5
5) Analytical approach	1	3	4	2	3.5	2	3	1.5	2	3	2	4.5	4	3	3	4	2	4.5	5	4
7) Activity planning and scheduling	2	3	2.5	4	3.5	4	4	3.5	4	3.5	2.5	3	4	3	5	5	5	3	6	5

Table 7: Distribution analysis: Summary of ratings applied to plans across the three lowest scoring areas

Key quality areas	Min	Max	Mean	Mode	Distribution	Less than satisfactory (0-3)	Satisfactory (4-6)
4) Strength of evidence	2.5	5	3.7	3		**45%**	55%
5) Analytical approach	1	5	3.1	3		**65%**	35%
7) Activity planning and scheduling	2	6	3.8	4		**50%**	50%

As previously mentioned, **the main area of weakness identified was the *analytical approach***. Only 35% (7 out of 20) of the plans scored 4 or higher, with an average score of 3.1 (see Box 1 for a good practice example). While most plans made some mention of data processing, the level of detail was generally limited. Only one plan fully addressed measures for data checking, error correction, secure storage and preparation for analysis, while the remainder provided minimal references to these considerations. In contrast, data analysis was covered in almost all plans but varied greatly in quality. **The weakest examples described analytical steps only in**

relation to the specific methods or datasets, without articulating how evidence from multiple sources would be integrated to answer the key evaluation questions (KEQs).

Consideration of disaggregated data analysis was also limited, appearing in only five plans and to a strong standard in just one. In many cases, this possibly reflected the nature of the evaluation rather than an omission, at least as far as primary data collection was concerned.[5] Similarly, only half of the plans described how evaluators would make judgments about the strength of evidence, and only a small minority explained how rubrics or criteria would be used to guide interpretation. Collectively, **these weaknesses suggested that, while most evaluators recognised the need for analysis, few demonstrated a systematic approach to transforming data into credible and transparent findings**. **This gap reduced the likelihood that evaluation results would be both robust and defensible, representing a key area for improvement in future planning.**

Box 1. Analytical approach – INL921 Kiribati Australia Skills for Employment Program (KASfEP) good practice example

The KASfEP Evaluation Plan describes a systematic analytical approach by outlining how quantitative and qualitative data will be processed, organised and analysed. Quantitative data, such as enrolments, completions, gender-disaggregated participation, and cost-per-participant metrics, are intended to be compiled in Excel and compared against program targets. Qualitative data from interviews, roundtable discussions, and questionnaires are to be coded using both deductive (KEQ-aligned) and inductive thematic analysis, supported by structured transcription and documentation processes. The plan also outlines the intention to analyse data disaggregated by gender, disability and other intersecting identity factors. Evaluative judgements are to be informed by clear criteria, triangulation across multiple data sources, and the use of an evaluation results matrix that links evidence to each KEQ and assesses its strength. Overall, the processes described establish a coherent pathway for transforming collected data into credible and transparent findings.

The second weakest area was the *activity planning and scheduling.* Only half (10 out of 20) of the plans scored 4 or higher, with an average score of 3.8 (see Box 2 for a good practice example). Approximately half of the plans clearly identified key respondents, preferred data collection methods or indicative visit locations, while the remainder provided only partial or limited coverage of these elements. Six plans had notable gaps, often omitting specific details on sequencing, such as the overall timeline of the evaluation. Time considerations were particularly underdeveloped, with limited reflection on data collection duration and content (e.g., the time allocated for interviews and the total number of questions), respondent numbers, or on how timing constraints might affect data quality. Only a quarter of the plans demonstrated strong planning in this regard, while the majority were regarded as operating under tight or unrealistic schedules. Overall, this indicated that, **although activity planning was generally recognised as a necessary component, its treatment tended to be cursory, reducing confidence in the feasibility and rigour of the planned data collection activities.**

Box 2. Activity planning and scheduling – INL921 Kiribati Australia Skills for Employment Program good practice example

The KASfEP Evaluation Plan provides a clear and practical activity schedule that aligns well with DMEL Standard 9.16. The plan identifies key respondent groups and preferred data collection approaches, including individual interviews, group discussions and questionnaires, and distinguishes between remote and in-country methods.

[5] The majority of evaluations reviewed did not collect primary quantitative data and as noted elsewhere, the quality and quantity of the secondary/MEL data available was often insufficiently described for the reader to understand the extent to which disaggregated data was available or not.

Annex 5 sets out a detailed stakeholder consultation plan that specifies the day, time, method, and location for each engagement, enabling DFAT to translate the indicative plan into a final schedule. The plan also makes clear where specific interviewees are still to be confirmed and incorporates flexibility to adjust the schedule as needed, including through a combination of paired teamwork, remote interviews, and backup consultation options. Together, these elements provide DFAT with sufficient information to plan meetings, negotiate access with stakeholders, and ensure adequate time allocation for meaningful data collection across all methods and respondent groups.

S*trength of evidence was* another area identified as a key weakness. Just over half (11 out of 20) of the plans scored 4 or higher, with an average score of 3.7 (see Box 3 for a good practice example). All plans described data collection methods that were broadly appropriate to the evaluation scope, though the level of elaboration varied considerably. Just over half provided adequate detail and linked methods back to the KEQs, while others remained high-level or descriptive. Interviews and document reviews were universally included as proposed methods, with some plans also featuring field observations and a smaller number incorporating surveys. One notable absence across several plans was the limited description of the monitoring and secondary data, and how this was to be used in the evaluation. It appears likely that this was considered as having been 'covered' by the document review; however, it warrants further attention in plans, given how integral such data is to most evaluations. Sampling strategies were presented in almost all plans, with the majority adopting purposive sampling approaches for qualitative methods, although, these were not always clearly justified or applied consistently across methods. Triangulation was mentioned in nearly all plans, yet few elaborated on how this would be operationalised, suggesting that, **while links between methods and questions were often clear, the process for integrating lines of evidence was underdeveloped.**

The consideration of respondents' needs, rights and security was more consistent, but uneven in depth. While most plans acknowledged these principles, fewer than half addressed them to a good standard. The stronger examples specified protocols for confidentiality, informed consent and data protection, alongside provisions to remove barriers to participation for disadvantaged groups. **Twelve plans annexed data collection tools or evaluative activities, though these were mixed in quality,** for example, several interview guides contained excessive or overlapping questions, and some survey instruments were poorly designed. The better examples included clear participant information, tailored question guides for different stakeholder groups, and thoughtful sequencing of questions to support data quality. Taken together, **the treatment of methods, sampling and ethics indicated that, while evaluators generally selected appropriate techniques, the rigour with which these were articulated and justified varied widely. This limited confidence that the planned evidence bases were sufficiently robust or transparent to support strong evaluative judgments.**

Box 3. Strength of evidence – INN855 Women's Fund Fiji (WFF) Mid-Term Review (MTR) good practice example

The WFF MTR Evaluation Plan provides good detail on how the evaluation is intended to generate evidence, outlining three complementary methods: document analysis, key informant interviews, and a Talanoa Survey, and describing how each contributes to the inquiry. Document review is described as shaping the design of semi-structured interviews by identifying issues, assumptions, and areas requiring further examination, demonstrating integration between methods, questions, and lines of evidence. The plan also sets out ethical procedures, such as verbal consent at interviews, opportunities for respondents to clarify evaluators' interpretations and the deliberate inclusion of women, people with disabilities, rural or remote organisations, and other socially disadvantaged groups. The approach to triangulation includes comparing findings across documents, stakeholder groups, and geographic levels, and involving non-grantees to test the validity of community-level insights. While the sampling rationale could be more explicit, the annexed stakeholder list illustrates an intention to engage a broad mix of actors across government, civil society, fund governance, and community groups to support a diverse evidence base.

Gender equality, disability and social inclusion (GEDSI) was evident across all plans but was uneven in scope and depth. One possible reason for this is that GEDSI is not a cross-cutting area of focus within DMEL Standard 9, appearing explicitly in only 4 out of the 19 criteria: limitations (9.4), evaluation questions (9.5), methods described for each question (9.8), and data collection methods are appropriate (9.9). While most plans acknowledged GEDSI as a thematic area, few embedded it systematically across methodological and operational elements. **This suggests that GEDSI integration is still treated primarily as a content consideration rather than a cross-cutting principle guiding the overall evaluation design.**

Strengths of the plans assessed

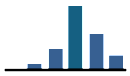
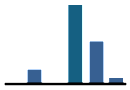
The three strongest areas across the evaluation plans, according to the average ratings assigned, were purpose and use of evaluation, roles and responsibilities and evaluation questions (Figure 4). These are elaborated further below. Notably, all three criteria were included in the plans' Terms of Reference (ToR). This meant that there was less elaboration required by the evaluators, unlike the other areas, which required more independent thinking, planning and elaboration. As above, these areas of strength also aligned closely with the strengths and weaknesses frequency counts (Annex 4).

Table 8: Heat map: Ratings applied to plans across the three highest scoring areas

This table uses red to green shading to provide a heat map overview of evaluation plan ratings, where low scores of 1 are dark red and strong scores of 6 are dark green. A version of this table summarising data distribution without colour shading is available in Annex 5, ***Table M****.*

Key quality areas	Plan 1	Plan 2	Plan 3	Plan 4	Plan 5	Plan 6	Plan 7	Plan 8	Plan 9	Plan 10	Plan 11	Plan 12	Plan 13	Plan 14	Plan 15	Plan 16	Plan 17	Plan 18	Plan 19	Plan 20
1) Purpose and use of evaluation	2	5	4	4	4	5.5	4	4.5	4	3.5	5.5	4	4	4	5	5	5	5.5	5	6
3) Evaluation questions	4	4.5	4	4	3	3.5	4	4	4	5	5	2.5	4	5.5	5	5	6	3.5	4	6
8) Roles and responsibilities	2	4	4	4	4	2.5	5	4.5	4	4	4	4	4	4	5	5	5	6	5	5

Table 9: Distribution analysis: Summary of ratings applied to plans across the three highest scoring areas

Key quality areas	Min	Max	Mean	Mode	Distribution	Less than satisfactory (0-3)	Satisfactory (4-6)
1) Purpose and use of evaluation	2	6	4.5	4		10%	**90%**
3) Evaluation questions	2.5	6	4.3	4		20%	**80%**
8) Roles and responsibilities	2	6	4.3	4		10%	**90%**

Purpose and use of evaluation **was the strongest area**, with 90% (18 out of 20) of the plans scoring a 4 or higher, with an average rating of 4.5 (see Box 4 for a good practice example). **All of the plans included a clear description of the program being evaluated**, which was usually positioned early in the document. This generally set out the rationale, scope and main features of the intervention. In most cases, this provided a solid foundation for understanding the focus of the evaluation and its logic. A small number of plans were more cursory, omitting detail on the operating context, end-of-program outcomes (EOPOs) or budget, though these gaps rarely obscured the overall intent. The purpose of the evaluation was well expressed in nearly all plans, aligning closely with the program description. They were typically framed around both results assessments and learning/recommendations for future work, with many distinguishing between primary and secondary objectives. Overall, there **was strong clarity of intent and evaluators generally understood what the evaluation was for and how findings would be used.**

The *identification of intended users and pathways for use* was also strong. Almost all the plans named their primary users, with many also noting secondary audiences and, in several cases, describing how findings would inform design, policy or decision-making. Approximately half mentioned plans for the publication of the final report, while others either did not mention it or indicated that DFAT approval would determine this. Some plans outlined measures to enhance accessibility, produce summaries, follow DFAT Standards or redact sensitive material. **Practices such as considering dissemination and stakeholder engagement aim to increase the overall utility of the evaluations and align them more closely with DFAT's learning and accountability objectives.**

Box 4. Purpose and use of evaluation – INM313 Australia Awards and Alumni Engagement Program – Philippines (AAAEP-P) good practice example

The AAAEP–P Evaluation Plan clearly outlines the program context, describing its duration, budget, delivery modalities, partner agencies and development objectives, alongside both historic and updated sets of EOPOs and IOs. The purpose of the evaluation is to assess whether AAAEP–P has delivered its intended outcomes and identify improvements for the remainder of the program and its successor investment. The plan explicitly identifies its primary users, including DFAT program managers, Philippine government partners, the managing contractor, and the future design team, and incorporates several mechanisms intended to enhance use, such as knowledge-sharing events, an Aide Mémoire webinar, and the engagement of an Evaluation Reference Group to guide the

feasibility and appropriateness of recommendations. The plan also documents DFAT's intent to publish the final report and states that it will adhere to DFAT accessibility requirements, supported by the involvement of DFAT's Development Evaluation and Assurance Section in providing quality assurance and advising on sensitivities before publication. These elements together show a clear and credible approach to ensuring that evaluation findings inform decision-making and are accessible to intended users.

***Roles and responsibilities* was the next strongest area.** Similarly to the above, 90% (18 out of 20) of the plans scored a 4 or higher, with an average rating of 4.3 (see Box 5 for a good practice example). The roles and responsibilities were clearly defined in most plans, with the majority providing sufficient detail for the evaluation team and outlining lines of accountability for key personnel. Approximately a quarter went further, delineating responsibilities across all parties involved, including DFAT and other stakeholders. Quality assurance (QA) processes were also well addressed, with roughly half of the plans rated as adequate and a further quarter demonstrating excellent practice. The strongest examples included defined QA mechanisms for all parties and a team member tasked specifically with reviewing outputs or providing feedback in a QA function. Only two plans neglected to mention QA altogether. Reference groups were rarely used, with only one plan establishing such a structure, though their role was well described. **Therefore, clear roles and accountability were a consistent strength across plans, but QA could be more systematically embedded.**

Box 5. Roles and responsibilities – INN302 - Mid Term Review of the South Asia Water Security Initiative (SAWASI) good practice example

The SAWASI Review Plan sets out clear roles and responsibilities across DFAT, implementers, partner governments, the review contractor and the review team. DFAT's responsibilities include commissioning the review, approving the ToR and Evaluation Plan, providing documents, arranging access to stakeholders, and giving feedback on emerging and draft findings. Implementers are tasked with facilitating access to information and supporting stakeholder engagement, while partner governments provide inputs through interviews and feedback.

Within the review team, individual responsibilities are well defined. The Team Leader is responsible for ensuring alignment with DFAT MEL standards, coordinating team inputs and leading analysis, while the Contractor Representative manages the contract and provides quality assurance of all review products. Quality assurance processes are reinforced through DFAT's review of the Evaluation Plan, discussion of preliminary findings in the Aide Mémoire session, and iterative feedback on the draft report. Together, these mechanisms provide structured oversight that supports rigour and credibility without compromising the independence of findings.

The final area of strength was the *evaluation questions*. In total, 80% (16 out of 20) of the plans scored a 4 or higher, with an average rating of 4.3 (see Box 6 for a good practice example). The majority of the plans included an appropriate number of KEQs – generally up to five, as per DFAT's DMEL Standards – providing good coverage of the program scope. However, almost three-quarters included an excessive number of sub-questions, which risked diluting the focus and overextending what could be realistically addressed in the final report. A small number also added further lines of enquiry, although the distinction between these and the sub-questions was not always clear. Less than half of the plans commented on the prioritisation of questions. Moreover, in some cases, the KEQs were poorly articulated, framed more as topic headings. **Collectively, though, evaluation questions were generally well-framed and aligned with the program's scope. However, too many sub-questions and limited prioritisation risked reducing focus and weakening the evaluability of some plans.**

All 20 plans integrated GEDSI considerations within their KEQs and sub-questions, with approximately one-third doing this well and the remainder adequately, but with scope for stronger integration across other question areas.

Box 6. Evaluation questions – INN855 Fiji Gender Country (FWCC) Program Plan good practice example

The FWCC Evaluation Plan presents a coherent set of four key evaluation questions, each supported by structured sub-questions that closely align with the KEQs and avoid unnecessary expansion of scope. The sub-questions are further refined through an additional column that sets out what DFAT wants to know, which deepens and clarifies each line of enquiry while maintaining a direct relationship to the KEQ. The plan includes examples of prioritisation, with greater depth directed to EOPO1 and EOPO2 where DFAT seeks the strongest evidence, and lighter treatment of EOPO3 and EOPO4 where DFAT has indicated satisfaction with current reporting. Gender equality and social inclusion are integrated throughout the KEQs, reflecting the nature of FWCC as a gender program and ensuring that all questions incorporate inquiry into gendered outcomes, inclusion of diverse groups and the responsiveness of services. Together, these elements demonstrate a well-structured question framework that supports DFAT's intended use of the evaluation by ensuring focus, relevance and analytic depth.

Further reflections on the quality of plans

Reviewers also noted **that plan length was an important factor**, with documents often being either overly detailed; making them hard to follow and digest; or too brief, leaving out essential information. While variability in plan formats was not inherently problematic, **key content was often dispersed across sections, making it difficult to quickly identify core elements.** Plans that closely followed the DMEL Standards in terms of format were generally easier to review and tended to receive more favourable ratings.

When evaluation questions are well-formulated, they often translate into effective data collection tools and more detailed planning. Robust analysis processes can lead to thorough triangulation and careful consideration of evaluative judgement. Similarly, good scheduling usually follows a well-structured methodology. Reflections on limitations also tend to be stronger when the methodology is sound. Additionally, feasibility and confidence in implementation quality is largely determined by planning components such as timelines, limitations, and roles and responsibilities.

While the methodology did not allow for an exploration of the influencing factors on quality (see the Limitations section), several **factors may have potentially influenced the quality of the plans**. Limited timing or insufficient time allocation can constrain the development of a plan. Similarly, being at an early stage in the process may limit the ability to elaborate the methodology fully. Familiarity with DFAT processes and experience in designing evaluations can also affect plan quality. Finally, familiarity with and adherence to the DMEL Standards appeared to be influential, with higher marks often observed when plans closely followed these guidelines.

9.3 Correlation between the quality of evaluation plans and the quality of final reports

There is limited correlation between the quality of a plan and the quality of the associated report at a macro level. High-quality reports are linked to low-, medium- and high-quality plans, as shown below:

- High-quality plans – 2 of 3 reports were rated as satisfactory
- Medium-quality plans – 1 of 3 were rated as satisfactory
- Low-quality plans – 1 of 3 were rated as satisfactory

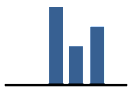
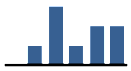
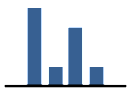
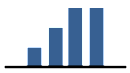
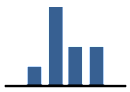
The three highest quality reports are based on evaluation plans from high-, medium- and low-quality categories (shown with an asterisk (*) and green font in **Table 10**).

Table 10: Heatmap: Report ratings by quality criteria, based on DMEL Standard 10, and comparison with plan ratings

This table uses red to green shading to provide a heat map overview of evaluation plan ratings, where scores of 1 are dark red and scores of 6 are dark green. A version of this table summarising data distribution without colour shading is available in Annex 5, **Table** *.*

Key quality areas [Reports]	L1 (Plan 1)	L2 (Plan 3)	L3 (Plan 5)	M1 (Plan 14)	M2 (Plan 16)	M3 (Plan 18)	H1 (Plan 17)	H2 (Plan 19)	H3 (Plan 20)
1) Purpose and use of evaluation	3.5	5	3	4	4.5	5.5	3.5	5	3
2) Evaluation design	3	4.5	2	2.5	2	4.5	3	5	5
3) Evaluation questions	3	4.5	2.5	3	5.5	5	3	6	6
4) Strength of evidence	3	4.5	3.5	3.5	5.5	5	4	5	4
5) Limitations	2.5	4.5	2.5	3	2.5	4.5	2.5	5	4
6) Recommendations and lessons	3	5	2.5	4	3.5	5.5	4.5	5	4
Overall report review score	**3.0**	**4.7***	**2.7**	**3.3**	**3.9**	**5.0***	**3.4**	**5.2***	**4.3**
Overall plan review score	2.1	3.4	3.6	4.1	4.5	4.6	4.6	4.9	5.4

Table 11: Distribution analysis: Summary of report ratings by quality criteria, based on DMEL Standard 10

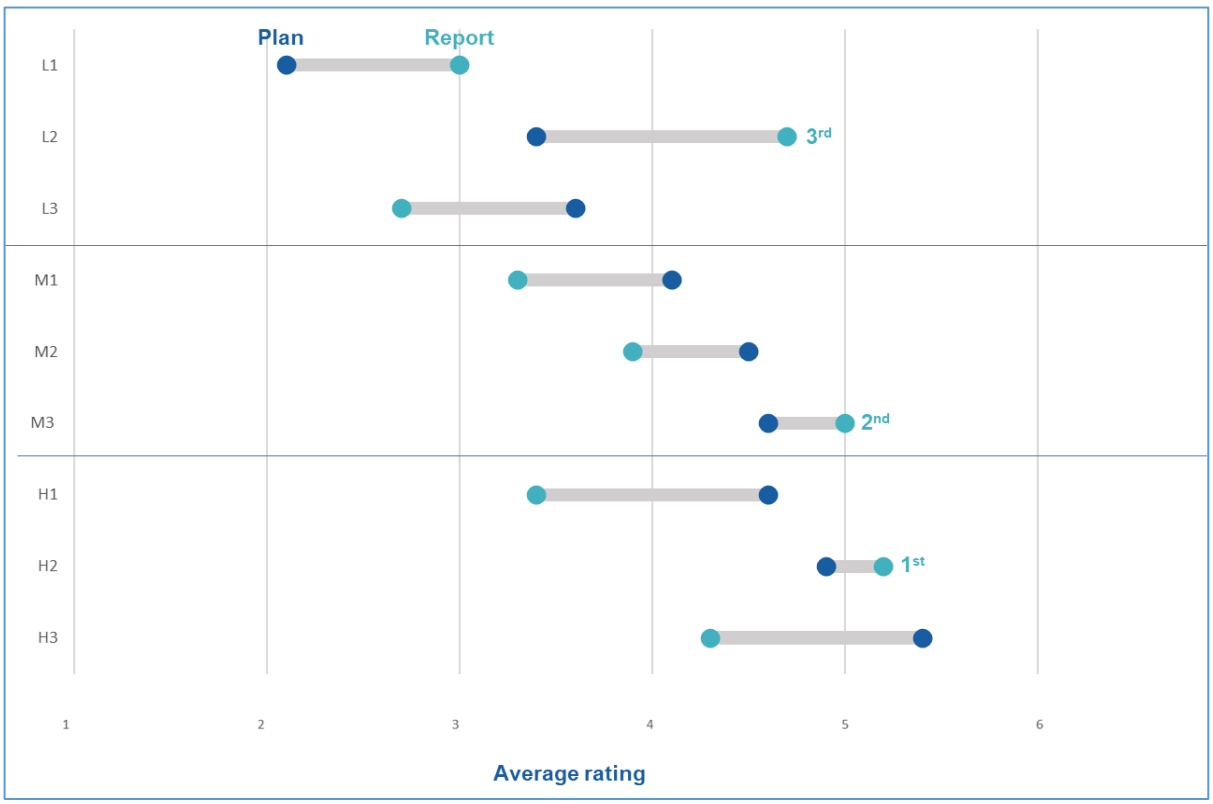
Key quality areas	Min	Max	Mean	Mode	Distribution	Less than satisfactory (0-3)	Satisfactory (4-6)
1) Purpose and use of evaluation	3	5.5	4.1	3		44%	56%
2) Evaluation design	2	5	3.5	2		56%	44%
3) Evaluation questions	2.5	6	4.3	3		44%	56%
4) Strength of evidence	3	5.5	4.2	3.5		33%	67%
5) Limitations	2.5	5	3.4	2.5		56%	44%
6) Recommendations and lessons	2.5	5.5	4.1	5		33%	67%
Overall report review score	**2.7**	**5.2**	-	-		**56%**	**44%**

*Having a high-quality plan did not guarantee a high-quality report. Quality ratings of reports were generally lower than plans across most quality areas (**Table 11**).* In five cases, the rating of the report was lower than the plan (as indicated in

Figure *2*) whereas in only four cases the average rating of the report was higher than the plan. This shows that planning clarity does not necessarily guarantee reporting rigour. Monitoring and follow-through during implementation are critical.

Figure 2: Average overall rating compared across plans and reports

*A screen-reader accessible version of the information in this figure is available in Annex 5, **Table Q**.*



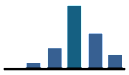
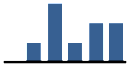
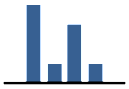
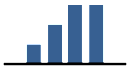
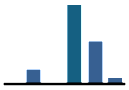
The team undertook more detailed analysis comparing criteria across the evaluation plan and report as shown in

Table 12.

Table 12: Plan ratings by quality criteria, based on DMEL Standard 10, and comparison with report ratings

An accessible version of the information in this table is available in Annex 5, **Table O and**

Plans (n=20)					**Reports (n=9)**				
Key quality areas	**Mean**	**Mode**	**Distribution**	**Satisfactory (rating 4-6)**	**Key quality areas**	**Mean**	**Mode**	**Distribution**	**Satisfactory (rating 4-6)**
1) Purpose and use of evaluation	4.5	4		**90%**	**1) Purpose and use of evaluation**	4.1	3		56%
2) Evaluation design	4.0	3		65%	**2) Evaluation design**	3.5	2		44%

3) Evaluation questions	4.3	4		**80%**	**3) Evaluation questions**	4.3	3		56%
4) Strength of evidence	3.7	3		55%	**4) Strength of evidence**	4.2	3.5		**67%**
6) Limitations	4.0	4		60%	**5) Limitations**	3.4	2.5		44%
-	-	-	-	-	**6) Recommendations and lessons**	4.1	5		**67%**
5) Analytical approach	3.1	3		35%	-	-	-	-	-
7) Activity planning and scheduling	3.8	4		50%	-	-	-	-	-
8) Roles and responsibilities	4.3	4		90%	-	-	-	-	-

This analysis showed that higher ratings for **purpose and use** in plans related to good definition of evaluation intent and governance. However, translating that clarity into robust evidence generation during implementation remains a gap, as indicated by the lower ratings for purpose and use in reports.

Evaluation questions were relatively stronger in plans (80%) compared with reports (56%), indicating that strong conceptual framing in plans sometimes weakened during implementation, negatively affecting the focus of analysis and clarity of reported findings.

Evaluation design and analytical approach were consistently weak in both plans and reports. Many plans lacked sufficient detail on data sources, sampling, and analytical methods, and reports demonstrated limitations in methodological rigour and evidence triangulation. This weakness in plans may be carrying through to reports, and limited upfront design quality can constrain evidence quality later.

Limitations were relatively weak across both plans (60%) and reports (44%), suggesting that while evaluators recognise the need to acknowledge limitations, these are not consistently being used effectively to contextualise the strength and reliability of findings.

More detailed analysis however suggests **a broad correlation between the quality of plan and the quality of the associated report**. This is shown by the increase in average ratings for the quality of the report associated with higher quality evaluation plans **Table 13**. Where plans

were of higher quality—both in meeting requirements and demonstrating strong analytical approaches—this was linked to greater depth of evidence in the reports.

Table 13: Ratings across quality categories, comparing plans and reports

Quality category	Average plan review score	Average report review score
Low	3.0	3.5
Medium	4.4	4.1
High	5.0	4.3

The **strength of evidence** improved from plan to report (55% to 67%), suggesting that some evaluations produced credible findings despite weaker design foundations, with implementation partially compensating for design gaps—though not consistently. Although most reports draw on a moderate evidence base, several recurring weaknesses were observed:

- Limitations in the quality or robustness of quantitative data (3 out of 9);
- Over-reliance on a single data source or method (2 out of 9); and
- Insufficient triangulation across data sources (4 out of 9).

9.4 Characteristics of good- and poor-quality evaluation plans

Table 14 presents the distinguishing characteristics of the highest and lowest rated plans, focused on high- and low-scoring quality criteria. It was noted that not all strong plans scored well in all quality areas, and the reverse was true for the weaker plans.

The good-quality plans tended to clearly articulate the purpose of the evaluation and how findings would be used. They included a focused set of evaluation questions, which reflected the program's objectives and incorporated GEDSI considerations. They provided sufficient methodological detail, including sampling strategies, analytical processes and annexed data collection tools, with appropriate use of monitoring and secondary data. They also outlined realistic schedules and clearly defined roles and responsibilities, including QA, while acknowledging limitations in the design or implementing an approach to mitigate their effect.

Conversely, poor-quality plans generally provided little detail on how the evaluation would be collaborative or utilisation-focused, with summaries that lacked key design elements and had limited consideration of ethics or flexibility. They included insufficient explanations of sampling strategies and made limited use of monitoring or secondary data, with data collection tools being underdeveloped or unavailable. Their analytical approach tended to be weak, with limited triangulation and no clear process for judging the strength of evidence, making it unclear how findings and conclusions would be reached.

Table 14: Characteristics of good- and poor-quality plans

Quality criteria (DMEL Standard 9)	Characteristics of good-quality plans	Characteristics of poor-quality plans
Purpose and use of evaluation	• Key program description elements were well summarised, including EOPOs, operating context and budget information. • Evaluation scope was specified.	• Lack of succinct summaries that capture key program detail. • Insufficient summary of context or scope. • Lack of articulation of purpose, users or intended usage.

Quality criteria (DMEL Standard 9)	Characteristics of good-quality plans	Characteristics of poor-quality plans
	• Relevant context was well described. • Intended users and usage were specified.	
Evaluation design	• Details of a collaborative approach were provided. • Overall design and high-level approach/methodology were explained and justified. • Potential ethical issues were well considered.	• Little description of a collaborative approach. • Utilisation focus not elaborated. • Methodology summary lacked key specifics or was not well explained. • Lack of flexibility woven into approach. • Ethics aspects were often light touch.
Evaluation questions	• Scope of questions reflected objectives. • A sensible number of KEQs and sub-questions were included. • GEDSI was considered either through specific questions or integrated into multiple questions.	• Scope of questions was not reflective of evaluation scope. • Number of evaluation questions or sub-questions were not appropriate (usually too many). • GEDSI was not well considered, often only seen as a brief add-on.
Strength of evidence	• Data collection methods were clearly specified and described. They were also appropriate for answering evaluation questions. • Sampling strategies were provided and sensible. • Data collection tools were of good quality and annexed. • Triangulation process was outlined. • Needs, rights and security of respondents were considered.	• Insufficient elaboration of data collection methods – they were not appropriate for effectively answering evaluation questions. • Insufficient tool development or availability. • No mention of monitoring or secondary data. • Little or no mention of sampling or triangulation processes. • Needs, rights and security of respondents given little or no attention.
Analytical approach	• Data processing and analytical approaches were described for all data collection components. • Disaggregation of data was discussed. • A process/approach for making judgements was outlined.	• Limited or no mention of data processing. • Analytical processes for data collection components were weak or not well elaborated. • Disaggregation was not discussed/ insufficiently discussed. • Process/approach for assessing strength of evidence was unclear.
Limitations	• Limitations/constraints were appropriate, sensible and effectively described. • Implications of limitations were discussed. • Mitigating measures were proposed.	• Limitations were limited in scope. • Implications or mitigating measures were either not included or weak.
Activity planning and scheduling	• Activity plans were well considered, with scheduling details, key respondents identified and indicative visit locations. • Time considerations were included.	• Limited accuracy in activity planning, such as inaccurate time estimations, inappropriate scheduling and a lack of feasibility. • Insufficient elaboration of key respondents.

Quality criteria (DMEL Standard 9)	Characteristics of good-quality plans	Characteristics of poor-quality plans
Roles and responsibilities	• Clear and sensible roles and responsibilities were included for all parties, including the review team members. • QA role was designated.	• Lack of clarity or information on roles and responsibilities. • QA function not included.

10. Recommendations to improve and strengthen evaluation plans

1. EVS to consider focusing on weaker elements in plans for additional support and training. These areas (listed below) could be elaborated on through existing tools and mechanisms. It is recommended EVS discuss the topics and create an Action Plan for implementation.

Collaborative approaches to evaluations

- What a collaborative approach entails and how it enhances the relevance, ownership and use of evaluation findings.

Key evaluation question (KEQs) and sub-question development

- Developing KEQs that focus on the most important aspects of performance.
- What prioritisation means in this context.
- Appropriate use and number of sub-questions.

Sampling

- Different sampling strategies suited to varying evaluation scopes.
- Defining a proportionate sampling approach within the evaluation plan.

Evidence and limitations

- Drawing on program monitoring, evaluation and learning (MEL) data as an evidence source.
- How to demonstrate varying strengths of evidence within evaluations.
- How to ensure collected data adequately answers the KEQs.
- Using an evaluation matrix (or equivalent) in evaluation planning for enabling linkage of questions to data sources.
- What triangulation means under the DFAT Standards.
- Benefits of clearly articulated limitations and aligning limitations, findings and conclusions to strengthen confidence in results and recommendations.

Analysis and making judgements

- Outlining information that should be included in evaluation plans to reasonably define approaches to data analysis and judgement-making.
- Proportionate approaches for different evaluation scopes.

Integration of GEDSI

- How GEDSI considerations can be embedded throughout the evaluation cycle, including within team composition, planning, question design and stakeholder engagement. This could be delivered jointly with DFAT's GEDSI team.

2. EVS should clarify and communicate expectations for DFAT staff and contractor use of DMEL Standards. This could be part of the Action Plan from Recommendation 1.

- Outline strategies to help clearly define and communicate DFAT's expectations for contractors' use of the DMEL Standards. This could include industry briefings.
- Support DFAT staff to apply the Standards when reviewing Terms of Reference (ToR), plans and reports.
- Communicate how DFAT staff can draw on existing support to reinforce these expectations.
- Work with DFAT procurement to further embed the Standards.

3. EVS to consider revising the DMEL Standards to improve clarity and to emphasise focus areas. Key standards noted for revision are outlined in **Table 15**.

Table 15: Recommendations for updates to the DMEL Standards

Quality criteria (DMEL Standard 9)	Recommendations for updates to the DMEL Standards
Standard 8.13, evaluation plan	The Standards suggest allocating three input days (depending on the program's scale and complexity) for developing the evaluation plan, including fully elaborated methods. Introducing a time range (i.e., 3-5 days) for input days rather than a fixed number of input days could encourage greater innovation in plan development, methodological design and overall evaluation conduct. This represents only a modest increase from current guidance. It is noted that time allocation will vary according to the complexity and scale of the investment. Further, commentary on the time duration for developing an evaluation plan should be included in the standards (e.g., 2-4 weeks). The duration may be longer to allow for a collaborative approach, especially when there are multiple stakeholders and to support localisation. The ToR could specify minimum expectations while highlighting areas where flexibility or innovation is encouraged. This would provide contractors with greater clarity and confidence when responding to tenders, helping them identify where they can add value beyond the ToR.
Standard 9.4, limitations	The Standards ask for limitations to be described, but this can result in a list of generic risks in the evaluation plan. The Standards should specify that limitations should be reflective of specific contextual program and evaluation issues/considerations rather than generic limitations. The Standards could more clearly distinguish between limitations (known shortcomings that impact on the scope and confidence in the evidence) and risks (potential challenges).
Standard 9.6, number of KEQs and prioritisation	The Standards emphasise having a small number of KEQs, however, this often leads to an excessive number of sub-questions in evaluation plans. The Standards would benefit from additional guidance on sub-questions. This would include what constitutes a reasonable number of sub-questions for the evaluation scope and to prioritise them to ensure the most relevant sub-questions are addressed.
Standard 9.8, data collection methods	Greater emphasis is needed on the role of the program's MEL system monitoring data. This could be viewed as a component of the document review, under 9.8. However, a clear sentence about the use of program MEL data and how the quality might be tested will support the data's inclusion in evaluation plans.
Standard 9.11, sampling	The Standard asks for a sampling strategy, appropriately allowing some flexibility according to the evaluation scope.

Quality criteria (DMEL Standard 9)	Recommendations for updates to the DMEL Standards
	Sampling guidance for quantitative and qualitative data collection could include justifications for sample sizes or clearer descriptions of approaches for purposive sampling in qualitative data. There could also be greater emphasis on protocols for reaching hard-to-reach groups (if relevant).
Standard 9.12, data processing and analysis	Guidance on data processing and analysis could be strengthened to encourage more focus in the evaluation plans. For example, the data analysis approach could describe the tools, processes (both individually and across the team), validation checks with stakeholders, and QA that will be used.
Standard 9.15, use of findings	The sub-heading could be expanded to include QA or the standard could be cross-referenced to Standard 9.18 (Quality Assurance). Activities such as testing early findings and initial findings workshops support use and are also data validation approaches that support QA.
Standard 9.16, schedule and Standard 9.7, flexibility	Standard 9.16 focuses on the schedule for data collection. The evaluation plan should include an overall schedule for the evaluation to ensure that adequate time is planned for key activities, including QA and DFAT/stakeholder review and feedback. This could include the indicative schedule for data collection that reflects the requirements of the standards. An overall schedule also supports understanding of the extent to which the evaluation can respond flexibly (Standard 9.7) within the timeframe. The guidance in Standard 8.14 (Scheduling in the ToR) is a useful reference for what to consider in the overall schedule.
Standard 9 – New – GEDSI integration	GEDSI aims in evaluation planning and conduct could be clearer in general. For example, are DFAT evaluations intended to be GEDSI sensitive or integrated?
Standard 10.4, limitations	The Standards should suggest the inclusion of a dedicated limitations section at the front of reports (e.g., in purpose and use). This will orient readers to the limitations of the findings and conclusions that follow and the level of confidence that can be placed in them. Additionally, the findings and conclusions section should be clear about the limitations of key assertions (where relevant) so that it is immediately clear for the reader/decision-maker (this may be linked to the assessment of strength of evidence, Standard 10.8).
Standard 10.8, strength of evidence	The Standards expect evaluations to clearly communicate the strength of evidence for key findings and related conclusions. As evidence is often synthesised in an evaluation, it is often not possible for a reader to determine the number of individual evidence sources, the use of triangulation or the implicit strength of evidence. Therefore, the standards should clarify how to demonstrate strength of evidence. That is, the evaluation should explicitly identify how it transparently comes to findings and conclusions. The Standard could also request the inclusion of a final, updated evaluation matrix (Standards 9.8, 9.10 and 9.14). This should demonstrate which type of evidence was used to answer each question and support an understanding of strength of evidence and triangulation.
10.19, recommendations: cost implications	The Standards ask for estimates of *"human, financial or material costs"*. Further advice on the expectations for cost estimations would support the application of the standard. For example, if a recommendation is to add a short-term adviser, do the costs of the role and any support costs need to be included?
10.20, lessons	The Standard asks that lessons should be specific to the program and the extent to which the lessons are transferable should be elaborated. The emphasis of the Standards is on the extent to which the lessons are transferable. However, they should more deeply emphasise the relevance of the lessons to the program (especially for a mid-term review or if there is likely to be a successor program) and the identification of the most critical lessons based on the findings.

List of Annexes

- **Annex 1**: Review plan
- **Annex 2**: Assessment template
- **Annex 3**: Summaries of pivot analysis of evaluation plans
- **Annex 4**: Frequency counts of strengths and weaknesses across plans
- **Annex 5**: Data tables with frequency of ratings

Annex 1: Review Plan

1. Introduction

The Development Evaluation and Assurance Section (EVS) of DFAT and Bluebird Consultants (Bluebird) will conduct a review of the quality of a sample of development evaluation plans and associated reports completed in 2024-25 (the review). This work follows five previous reviews of evaluation quality in 2012, 2014, 2017, 2023 and 2024. This document outlines the proposed method for the review to take place in 2025.

2. Background

The Australian Government committed to strengthened monitoring, evaluation, and learning approaches in its 2023 International Development Policy. This review provides data that contributes to one of the indicators of the Performance and Delivery Framework that underpins the policy. DFAT has committed to conducting an annual review of the quality and use of evaluations and publicly report on the findings as a measure of the indicator: *Our development cooperation is informed by monitoring, evaluation and learning*.

DFAT conducts around 40-50 program evaluations each year, in line with the Development Evaluation Policy ('the policy') introduced in 2016 and updated in 2023. Each of the former reviews of quality had a slightly different focus. The 2012, 2014 and 2017 reviews were undertaken by the former Office of Development Effectiveness (ODE). The 2014 ODE review examined the quality and credibility of evaluations as well as the factors influencing evaluation quality and utility. The 2017 ODE review focussed on assessing the impact of the revised Aid Evaluation Policy (2016) on evaluation practice, quality and use. The 2017 review also provided learnings from the evaluations on policy influence, aid capability and gender equality. The 2023 review examined the quality of evaluation reports and EVS conducted a separate analysis on the use of evaluation recommendations. The 2024 review assessed how well evaluation reports met quality criteria and how effectively evaluation findings were being used[6].

A key finding from the 2024 review was that DFAT staff lacked confidence in reviewing evaluation plans and reports in relation to DFAT DMEL Standards. DFAT's Development Effectiveness and Enabling Division (PRD) assumes that this relates to several factors, including uncertainty around responsibility for quality review of evaluation plans, varying levels of confidence in technically assuring evaluation plans and reports, and limited time available to conduct thorough quality assurance processes of plans and reports. Similar insights were also generated from DFAT staff supported through the Evaluation Helpdesk and Targeted Evaluation improvement Program[7]. EVS identified that it would be useful to focus the next review on the quality of evaluation plans and the extent to which, and how, this may influence the quality of evaluation reports.

3. Purpose

There are multiple purposes to this review. Primarily, the review will provide an opportunity to assess how well DFAT is implementing a key element of the Development Evaluation Policy –

[6] The 2023 and 2024 reviews were conducted by Bluebird.

[7] Two evaluation capacity building mechanisms supported through DFAT's EIS.

quality. In particular, it will explore common strengths and weaknesses in the evaluation plans reviewed, examine the correlation between the quality of the evaluation plan and the final report, identify characteristics of both strong and weak evaluation plans, and suggest practical actions to improve and strengthen future evaluation plans.

Secondly, the review will also test the assumption that a well-developed evaluation plan lays the foundation for a rigorous, credible, and useful evaluation, reflected in the quality of the final evaluation report.

From 2026 onwards, EVS is considering that quality assurance of evaluation plans in countries receiving intensive support through the Targeted Evaluation Improvement Program (TEIP) may become mandatory. If so, this review will also inform that process.

The review will assist in assessing the performance of Australia's international development program and will be published on DFAT's website.

The review will also be conducted in a way that builds EVS capability through close management of the work.

4. Objectives

The review has three objectives:

1. Contribute to measurement of the Tier 3 indicator in the Performance and Delivery Framework for the International Development Policy indicator *our development cooperation is informed by monitoring, evaluation and learning*.
    1. The measure is: *conduct an annual review of the quality and use of evaluations and publicly report on the findings*.
2. Support future assessment of progress and impact of the EIS, through establishing a baseline against which to assess support from EIS[8].
3. Track quality and use over time to inform DFAT policy and practice.

5. Scope

The review will be limited to "DFAT-led" evaluations (those commissioned and managed by a DFAT officer) completed in 2024-25 and key findings will be reported against the Tier 3 indicator in the next PADC report (2024-25).

The review will sample 20 evaluations from all 37 DFAT-led program evaluations completed between 1 July 2024-30 June 2025. The review will select a purposive sample of 20 investments enabling diversity and breadth (in line with investment emphasis) based on geographic coverage, value of investment and sector. The review will then sample 9 evaluation reports from the sample of 20, choosing 3 reports where the quality of the evaluation plan was high, 3 where quality was adequate, and 3 where quality was unsatisfactory or below.

[8] A baseline only applies to programs which have not yet received support from EIS.

Comparative review across both evaluation plans and reports will enable both a quality review of evaluation plans and an exploration of the linkage between evaluation plans and reports, as articulated through the review questions below.

6. Audience

The primary users of this review are staff from EVS, staff from the Design & Program Advice Section, senior managers in PRD, DFAT's Executive and the Development Program Committee. The findings from the review will be used by these groups to inform the department's approaches, policies, guidance, training and practices in MEL.

DFAT staff involved in commissioning and managing evaluations, performance and quality focal points, DFAT monitoring and evaluation advisers will have an interest in the review's findings in assisting them to better commission, support, manage, and use higher quality evaluations.

More broadly, the Australian public is also an audience, and the findings from the review will be published.

The findings of the review will be shared with DFAT staff and other stakeholders in the following ways:

- Key findings of the review will be published in the PADC Report.
- The full review report, including executive summary, will be published on the DFAT website.
- EVS staff and Bluebird (with DFAT's permission) will present the review process and findings at appropriate forums within DFAT and externally as opportunities arise. e.g., DFAT P&Q network, EvalNet meetings and with companies and consultants who undertake DFAT evaluations.

7. Review Questions

The review will seek to answer the following questions:

1. To what extent are DFAT evaluation plans good quality (meeting Standard 9)?
    a. What are the common areas of strengths and weaknesses of the evaluation plans assessed?
    b. What is the correlation between a quality evaluation plan and the quality of the final report?
2. What are the characteristics of good and poor-quality evaluation plans?
3. What practical actions are recommended to improve and strengthen evaluation plans (considering all tools available, including the EIS, the Standards and the Development Evaluation Policy)?

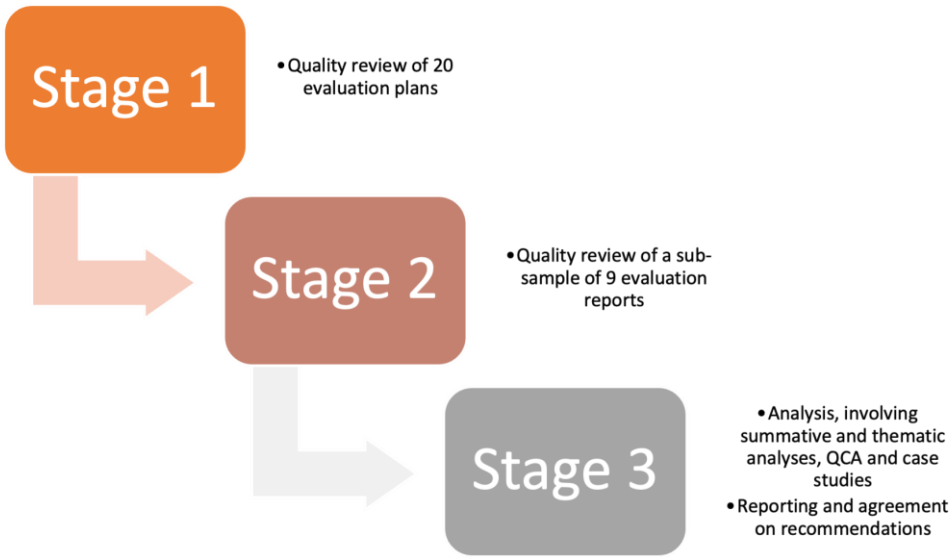
8. Methodology

The review will be conducted by a team of three external consultants. All three external consultants are fully independent, having had no involvement with the EIS core team or in delivering support services through the EIS, which positions them well to objectively assess the

plans and reports from EIS TEIP countries. The Team Leader was also involved in the 2024 review of quality and use and so will help to bring consistency to this review.

The review will be conducted over three stages, as outlined in Figure A below.

Figure A: Review stages



Stage 1 – Quality review of plans

The first stage will involve the sampling of all evaluations conducted during the review period to enable a purposive, broad sample of 20. The evaluation plans will be acquired from EVS, and the team will conduct a detailed quality review of each evaluation plan, against a tailored tool based on quality criteria linked to MEL Standard 9 – Independent Evaluation Plan. A moderation exercise will ensure consistency in approach across the team.

Stage 2 – Quality review of reports

The review team will then select a sub-sample of 9 plans to conduct a review of associated evaluation reports, ensuring correlation to a range of levels of quality in evaluation plans (three categories). This review will be conducted using a tailored tool linked to selected quality criteria based on Standard 10 - Independent Evaluation Report, which are relevant when comparing to MEL Standard 9.

Stage 3 – Analysis and reporting

The third and final stage will involve the team analysing the data to identify key findings and provide examples of varying quality of evaluation plans.

Specifically, the team will aggregate the ratings across the quality assessments of both the plans and the reports (separately) to explore the extent to which they are meeting Standard 9 and 10, respectively. These aggregate ratings will form a baseline. A thematic analysis approach will then be taken to explore and identify common areas of strengths and weaknesses of the evaluation plans, and then as a separate exercise, the evaluation reports.

To explore the correlation between the quality of evaluation plans and final reports, the team will conduct a QCA[9] using a sample of 9 evaluations for which both plans and reports have undergone quality review. A matrix will be developed to directly compare quantitative ratings and thematic analysis findings, organised according to the 3 quality categories. This will support consideration of the relationship between plan and report quality and inform the development of hypotheses for further exploration. This effort will be supported by the development of case studies of evaluation plans of high, moderate and low quality.

The team will collate preliminary findings and workshop these with the EVS review manager and Bluebird quality assurance focal person. The review team will then finalise the report, including developing a short snapshot of findings for senior executives, and communicate findings. Guided by EVS, the team could communicate findings at DFAT learning events (including for Evaluation Helpdesk Advisers and EIS Post Support Teams) and in other relevant internal and external forums.

Table A below summarises how each evaluation question will be answered.

Table A: Summary of methods by review question

Review question	Data collection methods	Data analysis methods
1. To what extent are DFAT evaluation plans good quality (meeting Standard 9)? • **What are the common areas of strengths and weaknesses of the evaluation plans assessed?** **What is the correlation between a quality evaluation and the quality of the final report?**	Basic characteristics of all 2024-2025 evaluations (and the investments they relate to) collected from the EVS management system and recorded in assessment template. A sub-sample of 20 evaluation plans rated against quality criteria in assessment template (based on MEL Standard 9) by expert evaluators. A sub-sample of 9 evaluation reports rated against quality criteria in assessment template (based on MEL Standard 10) by expert evaluators.	Summative (quantitative) and thematic (qualitative) review of evaluation plans. Summative (quantitative) and thematic (qualitative) review of evaluation reports. QCA to identify the correlation between the quality of evaluation plans and quality of associated evaluation reports.
2. What are the characteristics of good and poor-quality evaluation plans?	Data on quality of evaluation plans collected under Q1 above.	Development of case examples of evaluation plans of high, moderate and low quality.
3. What practical actions are recommended to improve and strengthen evaluation plans (considering all tools available, including the EIS, the Standards and the Development Evaluation Policy)?	Data on quality of evaluation plans and of a variable sample of associated reports collected under Q1-2 above.	Testing of preliminary recommendations. Finalisation of recommendations after EVS review of draft report.

The methodology relating to each review question is discussed in more detail below.

[9] For further information see: Simister, N. and Scholz, V., Qualitative Comparative Analysis, INTRAC for Civil Society, **https://www.intrac.org/wpcms/wp-content/uploads/2017/01/Qualitative-comparative-analysis.pdf**

1. To what extent are DFAT evaluation plans good quality (meeting Standard 9)? What are the common areas of strengths and weaknesses of the evaluation plans assessed? What is the correlation between a quality evaluation and the quality of the final report?

Review of evaluation plans

Sampling

As outlined above, the review will use a **purposive sample of 20 evaluation plans** from a wider sample of 37 DFAT-led evaluations completed between 1 January 2024-30 June 2025. The sampling approach will enable diversity and breadth in line with DFAT investment emphasis. Specific variables of interest include geographic coverage, sector, and value of investment. EIS support will not be considered as a variable to inform sampling given support so far received for evaluation planning via EIS has been minimal (applies to just one plan and two reports included in the sample)[10].

Results from this first round of sampling are captured in Table B below. A good range is acquired across both PADC categorisation (geographic region), sector and investment value[11].

Table B: Summary of sampled evaluations for quality of review of plans

PADC Categorisation	Sector	Investment Value
Pacific: 8 Southeast Asia: 5 South and Central Asia: 2 Beyond the Indo-Pacific: 1 Global: 1 Sector: 3	Governance: 9 Humanitarian: 1 Agriculture, Trade and other Production Sectors: 3 Education: 1 Health: 3 Economic Infrastructure and Services: 2 Cross-sector: 1	Low (<AUD$10m): 1 Medium (AUD$10–100m): 13 High (>AUD$100m): 6

Quality review process

The quality of each evaluation plan sampled will be assessed by the team. The tool for both the quality review of Plans and Reports is similar, though not all relevant quality criteria for MEL Standard 9 apply to MEL Standard 10, as indicated in the assessment template and Handbook. Given this is the first review to focus on Evaluation Plans, this tool has been developed specifically, though attempts have been made to use a tool similar in both format and scope as

[10] This is not considered sufficient to introduce bias into the baseline establishment process. It is also unclear at this point if the evaluations which received support for their reports will be included in the secondary sample.

[11] DFAT tend to consider below $10m as "low value" investments, which relates to thresholds for increasing quality requirements and levels of approval at the design stage (i.e. peer review, independent appraisal, Development Program Committee approval).

used in previous reviews to provide consistency, and enable some comparability and identification of trends as feasible.

The eight quality criteria are:

1. Purpose and use of evaluation
2. Evaluation design
3. Evaluation questions
4. Strength of evidence
5. Analytical approach
6. Limitations
7. Activity planning and scheduling
8. Roles and responsibilities

The team's assessment will be guided by a Handbook (Annex A) which includes greater detail on each of the eight criteria. The team comprises external (expert evaluators), with DFAT having a close oversight role, which means that the review will be informed by an inside understanding of key issues and priorities. The team will arrive at a rating of 1 to 6 (see Table C) and record a narrative comment in their rationale for assigning that rating, in the assessment template (Annex B).

Table C: Ratings

Rating	Satisfactory	Rating	Less than satisfactory
6	**Very high quality:** satisfies criteria in all areas.	3	**Less than adequate quality**: on balance, does not satisfy criteria and/or fails in at least one major area.
5	**Good quality:** satisfies criteria in almost all areas.	2	**Poor quality:** does not satisfy criteria in several major areas.
4	**Adequate quality**: on balance, satisfies criteria and does not fail in any major area.	1	**Very poor quality**: does not satisfy criteria in any major area.

Moderation

The ratings will be moderated in multiple ways, to increase the consistency of ratings across team members. First, the team leader was a member of the team from the 2024 Review of DFAT Evaluation Quality and Use, and this provides a level of continuity that is helpful in setting the bar for expectations of quality. Second, a moderation exercise will be conducted at the commencement of work. The exercise will involve the team members rating the same evaluation. This will ensure the team members are familiar with the template criteria, handbook and ratings criteria, ensure language, interpretations and relative priorities of aspects of the nine criteria are well understood across the team members. This moderation exercise will establish the basis for quality assessments. Following the moderation exercise, team members will undertake the quality reviews of all assigned evaluation plans. The completed assessment templates will be shared with and collated by the team leader for subsequent analysis.

Analysis

The team will **rate each plan on a scale of 1-6 against each core element of the tool**. The tool includes key categories, which aggregate the elements of Standard 9. These aggregate ratings will **be considered a baseline against which to consider EIS's impact** as the review predates the EIS.

A **thematic analysis approach will then be taken to explore and identify common areas of strengths and weaknesses of the evaluation plans**. These may be specific to quality criteria or cut-across quality criteria. These will be summarised, with (non-identifiable) examples, ahead of the next review phase.

Quality review of evaluation reports

Sampling

From the selection of evaluation plans, a **second round of purposive sampling** will be conducted to identify **9 evaluation reports** based on a breadth of quality in the evaluation plans as identified through the quality review. Average total review scores will be calculated for each evaluation plan reviewed before being organised into three categories, which relate to the ratings outlined in Table C above:

- High quality of evaluation plan – average score 5 and above
- Adequate quality of evaluation plan – average score 4
- Unsatisfactory or below quality of evaluation plan – average score 3 and below

Three plans will be selected from each category. Effort will also be made to ensure diversity (as far as possible) across the variables of interest which informed the original sample of 20 evaluation plans (i.e. geographic coverage, sector, value of investment and EIS support).

It is noted that the original intention had been to sample six evaluation reports in this second round of sampling (two from each of the above categories), but a larger sample is proposed based to enable wider evidence generation to inform possible linkage between quality of evaluation plans and reports. If this does not appear to be feasible according to time allocation, the review team will revert to the original plan to sample 6 evaluation reports.

Quality review process

The **sample of reports will be assessed against MEL Standard 10** using the same approach as outlined above for the review of evaluation plans. As with the evaluation plans, the team will arrive at a rating of 1 to 6 (see Table C) and record a narrative comment in their rationale for assigning that rating, in the assessment template (Annex B).

A moderation process is not required for this phase of the review given the same tool will be used for the quality review of plans and reports.

As per the review of evaluation plans, the team will **rate each report on a scale of 1-6 against each core element of the tool**, which includes key categories which aggregate relevant elements of MEL Standard 10. A **thematic analysis approach** will also then be taken to explore and identify common areas of strengths and weaknesses of the evaluation reports, to inform comparison with the evaluation plans.

To explore the correlation between a quality evaluation plan and the quality of the final report, the team will conduct a **QCA** based on the sample of 9 evaluations for which both plans and reports have been quality reviewed. A matrix will be developed to directly compare both the quantitative ratings as well as the thematic analysis findings, organised into the three quality categories outlined above. The results will be examined to explore and determine:

a) the degree of alignment in quality ratings between the evaluation plans and reports;

b) the extent and nature of shared strengths and weaknesses observed across both plans and reports, orientated around the quality criteria of MEL Standard 9; and

c) factor(s) which appear to be key in influencing the correlation between the quality of evaluation plans and reports. These may, for example, relate to achievement against quality criteria (i.e. well-defined scope, or solid methodology), DFAT role in the evaluation, or variables of interest in the sampling (i.e. investment value, geographic region).

These analytical foci will enable consideration of the correlation between the quality of evaluation plans and reports, as well as support the formulation of relevant hypotheses for further exploration.

2. What are the characteristics of good and poor-quality evaluation plans?

Drawing on the data collected under Question 1 above; **the identified characteristics of good and poor-quality evaluation plans will be profiled.**

Case examples of evaluation plans of high, moderate and low quality will also be developed – likely two from each category (a total of six). Good practice examples can be published on the DFAT website and shared with evaluation managers. Examples of poor practice will be drawn on to inform training and workshops focused on improving evaluation plans.

3. What practical actions are recommended to improve and strengthen evaluation plans (considering all tools available, including the EIS, the Standards and the Development Evaluation Policy)?

The team will develop recommendations drawing from the data and discussion with the broader EVS and Bluebird teams at a **collective workshop** held at the end of the analysis phase.

9. Roles and Responsibilities

The review will be undertaken by the following team members. The roles and responsibilities are noted below.

Team Leader:

- Draft the Review Plan and methodology for the review, update the template and manual
- Oversee review process for both evaluation plans and reports, including moderation/calibration workshops, lead the team, troubleshoot
- Moderate consistency of assessments across the team
- Conduct data analysis
- Draft and finalise the report

- Manage Team Members
- Liaise with Bluebird team and DFAT

Team Members:

- Contribute to the Review Plan as needed
- Assess the quality of sampled evaluations plans and reports by completing the assessment templates
- Conduct data analysis (quantitative and qualitative)
- Participate in moderation and weekly meetings and final workshop to share insights and help validate findings and propose recommendations
- Input into the draft and final report

Bluebird will ensure the quality of the final report through reviews by the Bluebird Director and quality assurance focal person. The final report will be copy edited and well-presented through the inputs of an editor and graphic designer. The report will meet accessibility requirements to be published on the DFAT website.

The DFAT review manager in EVS will provide close oversight and ongoing advice to the team throughout the review, provide feedback on the draft Review Plan and Review Report. DFAT will approve the final report.

10. Limitations and risks

i. Gauging overall evaluation quality

Assessing the quality of evaluation plans and reports is an important step in understanding overall evaluation quality. However, such assessments alone are insufficient without additional insights and evidence. This review does not include a broader documentation review or stakeholder interviews, both of which would strengthen the evidence base and the validity of conclusions on quality. While efforts were made to ensure diversity in the sample based on key variables, time and resource constraints meant that the final sample of evaluation plans and reports remains relatively small and is not fully representative of the broader set.

ii. Linkage between quality of evaluation plans and quality of evaluation reports

There is not necessarily a direct causal relationship between the quality of evaluation plans and the quality of evaluation reports. This review is based on underlying assumptions that need to be critically examined in order to generate meaningful insights and guide further reflection and learning in this area.

iii. Consistency of assessments across the team and across years

The quality of program evaluation plans and reports will be assessed by team members using the assessment template. To ensure the findings of the review are credible, it will be important to ensure team members assess evaluation plans relatively consistently to set a baseline against which future reviews and assessments of plans can be reasonably compared. Consistency across years is a risk, but the use of similar criteria and DFAT's DMEL Standards reduces this risk. Also, the team leader's involvement in the 2024 review helps minimise the

risk. Consistency of ratings across the team will be maximised by having a smaller team (than in prior years) and through the moderation processes described above.

iv. Managing sensitive information

The proposed approach raises potential risks and ethical challenges as team members may gain access to sensitive information through the document review and interview process. All team members will abide by DFAT's Ethical Research Guidelines and as professional evaluators are bound by the Australian Evaluation Society's Code of Ethical Conduct and Guidelines for the Ethical Conduct of Evaluations. All Bluebird team members have signed and are bound by confidentiality agreements.

v. Applicability of findings

There is a risk that findings gained only through the purposeful qualitative sample are generalised to all evaluations. The final report will be careful in presenting findings. The report will present findings that are triangulated across the analytical components.

vi. Conflict of interest

None of the Bluebird team members have been responsible for delivering TEIP or Helpdesk advisory services. As such, they have not had any involvement in improving the quality of evaluation plans or reports. All team members will declare any conflicts of interest, if they have been involved in any way with the programs that have been evaluated. If that is the case, another evaluator will be assigned to review the plan and report.

11. Schedule

The review will take place from July to October 2025, with the absolute deadline for the preliminary findings being Wednesday,15 October 2025. Table D below indicates the main tasks, persons responsible and approximate timeframe for the review tasks.

Table D: Review Tasks, Responsibilities and Timeline

Review Task	Person(s) responsible	Approximate dates for completion
Inception meeting to agree scope and approach of review plan	Team Leader and DFAT	July (Week 3)
Develop and finalise review plan	Team Leader	July (Week 3-4)
Conduct one evaluation review for moderation	Team Leader and Team Members	August (Week1-2)
Moderation workshop	Team Leader and Team Members	August (Week 2)
Review sample of 20 evaluation plans	All team members	August (Week 3-4)

Review Task	Person(s) responsible	Approximate dates for completion
Summation and analysis	Team Leader and Team Members	August (Week 4) September (Week 1)
Review sample of 9 evaluation reports	Team Leader and Team Members	September (Week 1)
Summation and analysis	Team Leader and Team Members	September (Week 2-3)
Collaboration analysis workshop	Team Leader and Team Members	September (Week 3)
Draft preliminary findings	Team Leader	September (Week 4)
Team workshop, including DFAT, to discuss analysis and agree on key findings and recommendations	All team members plus DFAT	October (Week 1)
Report drafting	Team Leader and Team Members	October (Week 2-3)
DFAT review of report	DFAT	October (Week 4)
Finalise report and summary outputs	Team Leader	November (Week 1-2)
Presentation of findings at a DFAT learning event	Team Leader and DFAT	November (Week 3-4)

11. Outputs

Outputs will include:

- This Review Plan outlining the detailed methods to be used for the review, including Assessment Template and Handbook and draft interview protocol.
- A concise report outlining the key findings of the quality review and recommendations, including:
    - A summary (such as small set of slides or a 2-page summary) for broadly communicating the findings
    - A list of 4-6 good practice evaluation products (annexed to the report)
    - Aggregated and de-identified lessons from poor practice use examples
- Detailed records of data collected.

Annex 2: Assessment template for analysis of plans and reports

Table E: Standard 9

Standard 9 Key quality areas and criteria	Standard 9 Quality Statements
1) Purpose and use of evaluation	▪ program description ▪ purpose of the evaluation ▪ primary intended users ▪ approaches to enhance use of findings ▪ plan for publication
2) Evaluation design	▪ a collaborative approach ▪ summary of the evaluation design ▪ flexibility to address important emerging/unexpected issues ▪ ethical issues
3) Evaluation questions	▪ key evaluation questions supplemented by detailed descriptions and/or sub-questions ▪ gender equality and social inclusion investigative approach
4) Strength of evidence	▪ methods are described for each question ▪ needs, rights, and security of respondents ▪ design of major evaluative/review activities/studies (with tools) ▪ triangulation of data collection methods ▪ sampling strategy
5) Analytical approach	▪ processing of data ▪ data analysis plan ▪ data disaggregation by sex and relevant socially disadvantaged groups ▪ process for making judgments
6) Limitations	▪ limitations or constraints effectively described ▪ implications are discussed
7) Activity planning and scheduling	▪ identification of key respondents, preferred data collection styles/methods and indicative visit locations ▪ inclusion of time considerations
8) Roles and responsibilities	▪ roles and responsibilities of team members, DFAT and any reference group ▪ quality assurance
Overall comments	[3-4 key, overall reflections]
High quality components	[List key aspects of high quality]
Low quality components	[List key aspects of low quality]
Should the Plan be considered for a case example?	Not applicable
Average quality score	[Score]

Table F: Standard 10

Standard 10 Key quality areas and criteria	Standard 10 Quality Statements
1) Purpose and use of evaluation	▪ purpose of evaluation ▪ executive summary ▪ resource implications of recommendations
2) Evaluation design	▪ summary of the methods employed (with elaboration in annexes)
3) Evaluation questions	▪ key evaluation questions applied with ease of navigation ▪ gaps explained
4) Strength of evidence	▪ clarity of author position, with unambiguous professional judgments ▪ line of sight from the key evaluation questions to evidence presented ▪ strength of evidence supports conclusions and judgments ▪ clarity of priority issues ▪ balance between operational and strategic issues ▪ role of context and emergent risks ▪ robust evidence and neutral language ▪ implications of key findings
5) Limitations	▪ key limitations enable appropriate interpretation of findings ▪ specific guidance regarding where caution needed
6) Recommendations and lessons	▪ feasible recommendations with responsibility allocated ▪ cost implications of recommendations ▪ transferable lessons
Overall comments	[3-4 key, overall reflections]
High quality components	[List key aspects of high quality]
Low quality components	[List key aspects of low quality]
Average quality score	[Score]

Annex 3: Evaluation plans: pivot analysis

Table G: Average quality scores across plans by geographic coverage (PADC)

Average quality rating	Pacific	Southeast Asia	South and Central Asia	Beyond the Indo-Pacific	Global	Sector	Total
Unsatisfactory	4	3	0	1	1	3	12
Adequate	4	1	2	0	0	0	7
High quality	0	1	0	0	0	0	1

Table H: Average quality scores across plans by sector

Average quality rating	Agriculture, trade and other production sectors	Economic infrastructure and services	Education	Governance	Health	Humanitarian	Cross-sector	Total
Unsatisfactory	1	0	0	7	2	1	1	12
Adequate	1	2	1	2	1	0	0	7
High quality	0	0	1	0	0	0	0	1
Total	**2**	**2**	**2**	**9**	**3**	**1**	**1**	**20**

Table I: Average quality scores across plans by investment value

Average quality rating	Low (<$10m)	Medium ($10–100M)	High (>$100M)	Total
Unsatisfactory	0	7	5	12
Adequate	1	5	1	7
High quality	0	1	0	1
Total	**1**	**13**	**6**	**20**

Annex 4: Frequency count of strengths and weaknesses across plans

Table J below outlines the frequency counts of strengths and weaknesses (up to 5 each) identified by reviewers across plans, focused on quality sub criteria (focus areas of interest in DMEL Standard 9 within each quality criteria).

The analysis statement (last column) highlights that program description, key evaluation questions, and roles and responsibilities are identified as areas of particular strength based on a high number of counts (above 5). However, key evaluation questions is also an area which received a number of counts as an area of weakness, suggesting a mixed picture here. Other quality sub criteria which receive a solid number of counts for strength as well as weakness include data collection methods, and limitations. Further to these, areas of particular weakness also include ethical issues, annexes of major evaluative activities, studies or tools, analysis of data disaggregated by sex and socially disadvantaged groups, process for making judgements, and time considerations.

Overall, this is broadly consistent with the findings from the average ratings by quality criteria discussed above as well as the qualitative analysis, but this approach enables elaboration of the quality sub criteria which drove the average ratings. It is also important to note the areas with consistently lower counts: while these did not emerge as clear weaknesses, they likewise did not present as strengths, indicating scope for improvement in these quality sub criteria. It is also noted that while reviewers were able to select up to five quality sub criteria as a strength of weakness for each plan, the did not specifically select five of each for every plan. In terms of overall counts, reviewers selected more weaknesses than strengths based on what was notably apparent – a count of 64 strengths as compared with 83 weaknesses.

Table J: Frequency counts of strengths and weaknesses across plans

Quality criteria	Quality sub areas	Count of strengths	Count of weaknesses	Symbol	Analysis statement
1) Purpose and use	Program description	8	1	▲	a strength across many plans
1) Purpose and use	Purpose	2	1	-	Not identified as an area of strength or weakness
1) Purpose and use	Primary intended users	3	0	-	Not identified as an area of strength or weakness
1) Purpose and use	Approaches to enhance use of findings	1	0	-	Not identified as an area of strength or weakness
1) Purpose and use	Plan for publication	2	1	-	Not identified as an area of strength or weakness
2) Evaluation design	Collaborative approach	4	1	▲	a strength across some plans
2) Evaluation design	Summary of evaluation design	3	1	-	Not identified as an area of strength or weakness
2) Evaluation design	Flexibility	2	0	-	Not identified as an area of strength or weakness
2) Evaluation design	Ethical issues	3	9	▼	a weakness across many plans
3) Evaluation questions	Key evaluation questions	5	5	▲/▼	a strength in some plans; a weakness in others
3) Evaluation questions	GESI investigative approach	1	0	-	Not identified as an area of strength or weakness
4) Strength of evidence	Data collection methods	4	9	▲/▼	a strength in some plans; a weakness in many plans
4) Strength of evidence	Needs, rights, and security of respondents	4	2	▲	a strength across some plans
4) Strength of evidence	Annexes: major evaluative activities, studies, tools	0	5	▼	a weakness across some plans
4) Strength of evidence	Triangulation	2	1	-	Not identified as an area of strength or weakness
4) Strength of evidence	Sampling strategy	1	2	-	Not identified as an area of strength or weakness
5) Analytical approach	Data processing	1	4	▼	a weakness across some plans
5) Analytical approach	Data analysis	2	9	▼	a weakness across many plans
5) Analytical approach	Analysis of data disaggregated by sex, socially disadvantaged groups	0	6	▼	a weakness across many plans

Quality criteria	Quality sub areas	Count of strengths	Count of weaknesses	Symbol	Analysis statement
5) Analytical approach	Process for making judgments	0	7	▼	a weakness across many plans
6) Limitations	Limitations	4	5	▲/▼	a strength in some plans; a weakness in others
6) Limitations	Implications	0	0	-	Not identified as an area of strength or weakness
7) Activity planning and scheduling	Key respondents, preferred data collection styles/methods and visit locations	2	4	▼	a weakness across some plans
7) Activity planning and scheduling	Time considerations	0	6	▼	a weakness across some plans
8) Roles and responsibilities	Roles and responsibilities (incl. independence)	7	2	▲	a strength across many plans
8) Roles and responsibilities	Reference group	1	0	-	Not identified as an area of strength or weakness
8) Roles and responsibilities	Quality assurance	2	2	-	Not identified as an area of strength or weakness
Totals	**Totals**	**64**	**83**	No data	No data

Key: ▲ Strength; ▼ Weakness; – Neither a strength nor a weakness.

Annex 5: Data tables with frequency of ratings

Throughout this report, bar charts nested within tables show how plan and report scores were distributed across quality areas. The following tables present that data in a screen reader accessible format.

Table K: Data to accompany Table 4: Heat map: Plan ratings by quality criteria, based on DMEL Standard 9

Key quality areas [Plans]	Count of 1s	Count of 2s	Count of 3s	Count of 4s	Count of 5s	Count of 6s
1) Purpose and use of evaluation	0	1	1	9	8	1
2) Evaluation design	0	3	4	8	3	2
3) Evaluation questions	0	1	3	9	5	2
4) Strength of evidence	0	1	8	8	3	0
5) Analytical approach	2	5	6	6	1	0
6) Limitations	1	0	7	7	2	3
7) Activity planning and scheduling	0	3	7	5	4	1
8) Roles and responsibilities	0	2	0	11	6	1
Overall plan review score	**0**	**1**	**11**	**7**	**1**	**0**

Table L: Data to accompany Table 6: Heat map: Ratings applied to plans across the three lowest scoring areas

Key quality areas [Plans]	Count of 1s	Count of 2s	Count of 3s	Count of 4s	Count of 5s	Count of 6s
4) Strength of evidence	0	1	8	8	3	0
5) Analytical approach	2	5	6	6	1	0
7) Activity planning and scheduling	0	3	7	5	4	1

Table M: Data to accompany Table 8: Heat map: Ratings applied to plans across the three highest scoring areas

Key quality areas [Plans]	Count of 1s	Count of 2s	Count of 3s	Count of 4s	Count of 5s	Count of 6s
1) Purpose and use of evaluation	0	1	1	9	8	1
3) Evaluation questions	0	1	3	9	5	2
8) Roles and responsibilities	0	2	0	11	6	1

Table N: Data to accompany Table 10: Heatmap: Report ratings by quality criteria, based on DMEL Standard 10, and comparison with plan ratings

Table 11: Distribution analysis: Summary of report ratings by quality criteria, based on DMEL Standard 10

Key quality areas [Reports]	Count of 1s	Count of 2s	Count of 3s	Count of 4s	Count of 5s	Count of 6s
1) Purpose and use of evaluation	0	0	4	2	3	0
2) Evaluation design	0	3	2	2	2	0
3) Evaluation questions	0	1	3	1	2	2
4) Strength of evidence	0	0	3	3	3	0
5) Limitations	0	4	1	3	1	0
6) Recommendations and lessons	0	1	2	3	3	0
Overall report review score	**0**	**1**	**4**	**2**	**2**	**0**

Table O and P: Data to accompany Table 12: Plan ratings by quality criteria, based on DMEL Standard 10, and comparison with report ratings

Key quality areas [Plans n=20]	Count of 1s	Count of 2s	Count of 3s	Count of 4s	Count of 5s	Count of 6s
1) Purpose and use of evaluation	0	1	1	9	8	1
2) Evaluation design	0	3	4	8	3	2
3) Evaluation questions	0	1	3	9	5	2
4) Strength of evidence	0	1	8	8	3	0
5) Analytical approach	2	5	6	6	1	0
6) Limitations	1	0	7	7	2	3
7) Activity planning and scheduling	0	3	7	5	4	1
8) Roles and responsibilities	0	2	0	11	6	1
Overall plan review score	**0**	**1**	**11**	**7**	**1**	**0**

Key quality areas [Reports n=9]	Count of 1s	Count of 2s	Count of 3s	Count of 4s	Count of 5s	Count of 6s
1) Purpose and use of evaluation	0	0	4	2	3	0
2) Evaluation design	0	3	2	2	2	0
3) Evaluation questions	0	1	3	1	2	2
4) Strength of evidence	0	0	3	3	3	0
5) Limitations	0	4	1	3	1	0
6) Recommendations and lessons	0	1	2	3	3	0

Key quality areas [Reports n=9]	Count of 1s	Count of 2s	Count of 3s	Count of 4s	Count of 5s	Count of 6s
Overall report review score	**0**	**1**	**4**	**2**	**2**	**0**

Table Q: Data to accompany Figure 2: Average overall rating compared across plans and reports

Case	Plan	Report	Difference
L1	2.1	3	0.9
L2	3.4	4.7	1.3
L3	3.6	2.7	-0.9
M1	4.1	3.3	-0.8
M2	4.5	3.9	-0.6
M3	4.6	5	0.4
H1	4.6	3.4	-1.2
H2	4.9	5.2	0.3
H3	5.4	4.3	-1.1