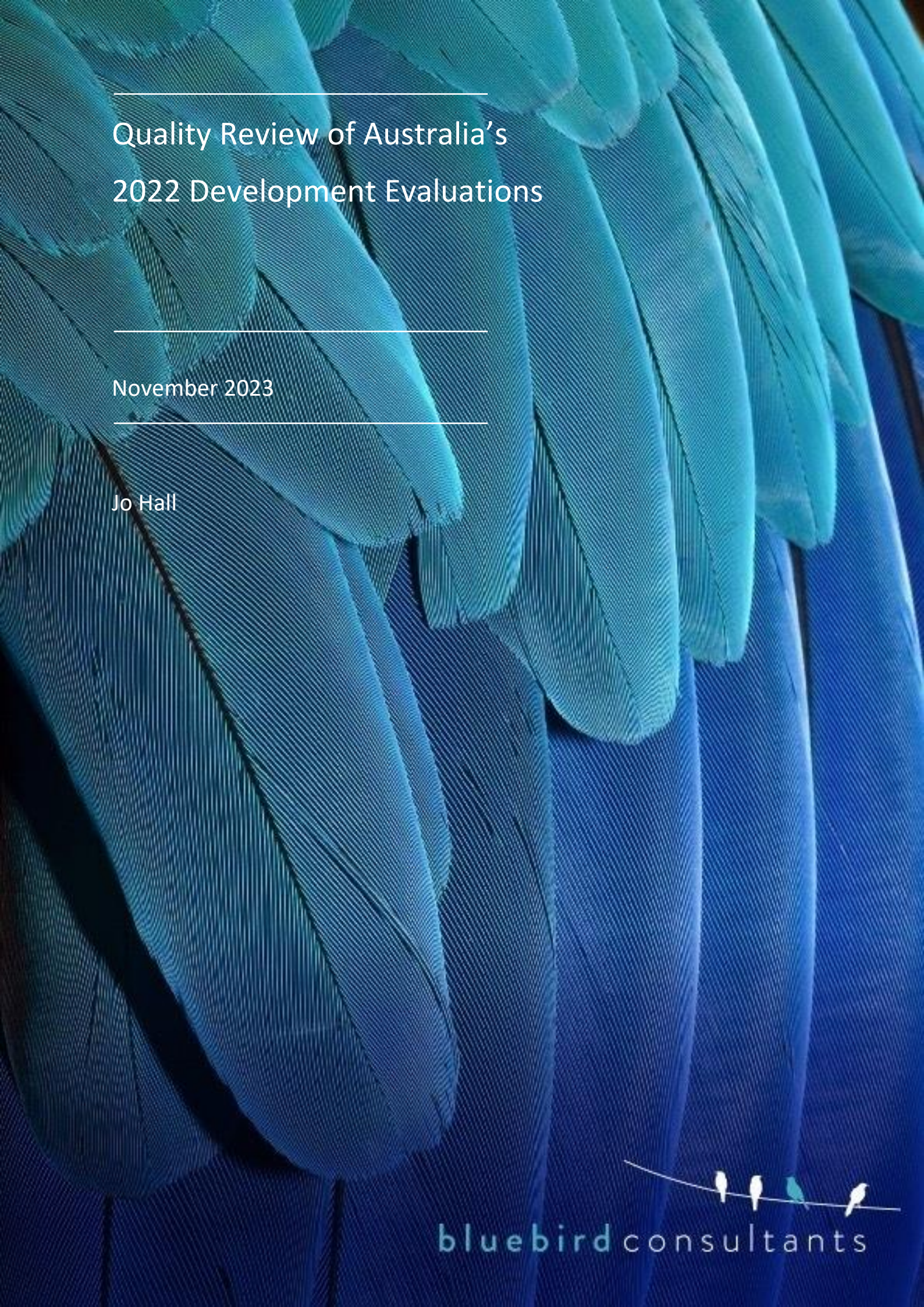
Quality Review of Australia's 2022 Development Evaluations

November 2023

Jo Hall



bluebird consultants

EXECUTIVE SUMMARY

Evaluation is strongly emphasised in DFAT's new Performance and Delivery Framework for Australia's International Development Policy. To that end, DFAT has committed to reviewing the quality and use of evaluations completed each year.

This report provides an assessment of the quality of all 40 DFAT development evaluations completed in 2022. Each evaluation report was analysed against nine quality criteria based on DFAT's Design and Monitoring and Evaluation Standards, using a similar approach to previous reviews. A separate report authored by DFAT examines the use of evaluations.

The findings from this quality review fall into three main areas:

1. **Most evaluations (70%) provide a credible source of evidence for the development program, but there is room for improvement.**

Twenty-eight (70%) of the 2022 evaluations were rated adequate or better (rated '4' or higher on a six-point scale). Almost half (17) of the evaluations were good or high quality (rated 5 or 6).

There is room for improvement as twelve evaluations were rated less than adequate quality. There was a wide range in the quality of evaluations across the Australian development program with no discernible pattern. This suggests some inconsistency in terms of good or poor practice.

2. **Quality of evaluations has declined slightly since 2012. There has been no meaningful change in overall quality since 2017.**

While overall quality has not changed over recent years, some evaluations are demonstrating improvement against particular quality criteria. DFAT's evaluations are addressing evaluation questions with increasing rigour, and the quality of executive summaries has improved over time. However, the appropriateness of methodologies and use of sources in evaluations have not improved in recent years.

3. **The 2022 evaluations include several standout performers, with three reports rated very high quality. By contrast, there were none in 2017.**

DFAT-led evaluations were slightly higher quality on average than partner-led or joint evaluations. Smaller investments (less than $10m) tended to have poorer quality evaluations.

The review considered the implications of these findings for DFAT as it moves to strengthen evaluation in the agency, as well as options for how DFAT's evaluation quality might be assessed in efficient annual processes. There is an opportunity to look for more timely and consistent assessments of evaluations earlier in the process, including at Terms of Reference and evaluation plan stages. This would serve two purposes: as a real-time feedback loop to strengthen and improve the evaluation as it progresses and as an ongoing data collection exercise contributing to the annual review.

The focus of this review is to assess how well the evaluation reports demonstrate the quality criteria for good evaluation practice. But there are other perspectives on quality and use of evaluations, for example, the quality of the evaluation *process* is another consideration, and can have more influence on practice by program and partner staff than the written report. The opportunity to take a broader perspective on quality and use of evaluations and modifying the methodology for the future annual review is discussed.

To address the findings of this review and options for future reviews, it is recommended that the Development Risk, Implementation and Evaluation Branch in DFAT:

1. **Continue to support DFAT program areas to manage high quality evaluations, with a particular focus on:**
    a. Meeting DFAT's Design and M&E Standards such as by formulating relevant evaluation questions, having an appropriate methodology to answer them, ensuring the evaluation report clearly addresses each question and provides sufficient evidence to support its findings.
    b. Sharing good practice examples.
    c. Promoting and building DFAT staff capacity to actively manage evaluations.

2. **Coordinate early review of evaluation TORs, evaluation plans and draft reports by expert providers to improve evaluation quality in a timely manner and integrate these assessments with the annual review.**
    a. This could be done by enhancing the work of existing quality assurance panels already contracted to program areas and supporting DFAT's Development Evaluation and Assurance Section with additional expert providers as needed.
    b. These expert providers could promote consistency of assessments, support DFAT's capacity building efforts, and provide data for the annual review.

3. **Articulate a broader perspective on quality and use of evaluations and define the scope of each annual review within this framework.**
    a. Adapt the methodology for the annual review of the quality of DFAT evaluations to match the revised scope, for example include interviews with staff and partners to capture a fuller understanding of evaluation quality beyond the report assessment.
    b. Further update and streamline the methodology by revising the evaluation criteria, assessment template and handbook to meet the latest DFAT Design and M&E Standards and to address some of the methodological limitations of this review and learnings from implementation. For example, reduce the number of criteria to approximately 5 key areas and ensure the sub-elements within the criteria are not duplicated.

ACKNOWLEDGEMENTS

This review was undertaken by Bluebird Consultants, with team leader Jo Hall and team members Kari Sann, Anna Roche and Farida Fleming. Bluebird Director Jess Kenway provided oversight. Penny Nettlefold, Assistant Director of Development Evaluation and Assurance Section in the Development Risk, Implementation and Evaluation Branch of the Department of Australian Affairs and Trade (DFAT) was also a valued team member. We are grateful to other members of DFAT's Development Evaluation and Assurance Section for information provided, oversight and other assistance in preparing this review. And of course, to the authors of the 40 evaluation reports assessed for this review.

CONTENTS

BACKGROUND

DFAT contracted Bluebird Consultants to conduct the Quality Review of Development Evaluations (the review). This review examines the quality of all DFAT development program evaluations completed in 2022, assessing the degree to which the published reports reflected DFAT's quality criteria for good evaluation practice.[1] It follows three previous reviews of program evaluations, which examined evaluations completed in 2012, 2014 and 2017. While these reviews used the same basic methods (such as ratings by an expert team against the same nine quality criteria), there have been significant differences:

- The first review used contracted consultants to review all 87 independent program evaluations completed in 2012.
- The second review was conducted by DFAT staff, who reviewed 35 program evaluations. This was a purposeful sample from the 77 program evaluations completed in 2014.
- The third review examined all 37 program evaluations identified in the Aid Evaluation Plan and completed in 2017. It was conducted by DFAT staff and one contracted consultant.
- The current fourth review examined all 40 program evaluations identified in the Aid Evaluation Plan and completed in 2022. It was conducted by four contracted consultants plus a DFAT staff member.

DFAT also completed reviews of the use of program evaluations in 2014, 2017 and 2021, sometimes in concert with the review of the quality of evaluations.

PURPOSE

DFAT's Performance and Delivery Framework supports the implementation of Australia's International Development Policy, which was released in August 2023. It includes a strengthened approach to evaluation emphasising the quality and use of evaluations to guide decisions on development programming. Performance is assessed using the indicator "Our development cooperation is informed by monitoring, evaluation and learning" [2]. One of the measures for the indicator is to "Conduct an annual review of the quality and use of evaluations and publicly report on the findings". This report presents findings on the quality of evaluations and will be published. A separate DFAT authored report looks at the use of evaluations. Together, they enable DFAT to understand the current status of evaluation quality and use and to report on the measure.

The review has three objectives:

1. To better understand the practices related to and quality of independent program evaluations, and how these have changed over time against findings from similar reviews conducted in 2012, 2014 and 2017.
2. To inform approaches to strengthen evaluation across the department.
3. To provide evidence that Australia's development cooperation is informed by monitoring, evaluation and learning.

[1] The Review assessed the evaluation reports against a combined set of criteria drawn from DFAT's 2017 Monitoring and Evaluation Standards (see Annex 3). Other perspectives on the quality and use of evaluations are considered in the Discussion chapter.

[2] Australia's International Development Performance and Delivery Framework, Department of Foreign Affairs and Trade, August 2023, page 7.

METHODOLOGY

A team of four expert evaluators from Bluebird Consultants and a DFAT Development Evaluation and Assurance Section staff member assessed the quality of all 40 evaluations completed in 2022 (Annex 2) by reviewing the evaluation reports against the nine quality criteria used in previous years. These criteria are as follows:

1. Purpose of evaluation
2. Scope of evaluation
3. Appropriateness of methodology and use of sources
4. Adequacy and use of M&E
5. Context of the initiative
6. Evaluation questions
7. Credibility of evidence and analysis
8. Recommendations
9. Executive summary

Guided by a handbook (Annex 3), the team members arrived at a rating of 1 to 6 (see Table 1) for each criterion and recorded narrative comments on their rationale for assigning that rating in an Excel template. See the full methodology in the Review Plan at Annex 1.

Table 1: Ratings

Satisfactory		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; does not fail in any major area	1	**Very poor quality:** does not satisfy criteria in any major area

LIMITATIONS

1. **Consistency of assessments across the team and across years.**

- To ensure the findings of the review were credible, it was important that team members assessed program evaluations as consistently as possible. Ratings were moderated across the team in several ways, including a moderation exercise involving all team members assessing the same evaluation, weekly team meetings to discuss application of particular criteria and oversight from the team leader (see Annex 1 for details). The team leader, for example, conducted a basic check for consistency between the narrative comments and the ratings as each assessment was completed, requesting a review by the team member in cases of apparent inconsistency.
- It was also important that this assessment could be reasonably compared to assessments of 2012, 2014 and 2017 evaluations. The use of the same nine criteria and DFAT's Design,

Monitoring and Evaluation Standards (2017) reduced the risk of inconsistencies. These risks were further reduced with a smaller team than in prior years and the team leader's involvement in the 2012 and 2017 reviews.

2. Methodological limitations

- The nine quality criteria have multiple sub-criteria, not all carrying equal weight in the assessment. For example, 'appropriateness of methodology and use of sources' includes the methods used; the limitations of the evaluation; and how ethical issues were addressed. The methods used carried more weight in the assessment than the other two aspects. There was some duplication across criteria, such as methodology, which was assessed in both 'scope' and 'appropriateness of methodology and use of sources'. This complexity was addressed through the weekly team meetings to ensure team members applied a consistent approach.
- If information such as the Terms of Reference were missing, it proved harder to rate reports for some criteria, especially scope and purpose, and, to lesser degrees, appropriateness of methodology and use of sources and the context of the initiative. As a result, some ratings for those criteria inevitably included a measure of compliance as much as quality. TOR and evaluation plans were not reviewed unless they were attached to the evaluation reports as an annex.
- The 2014 review was of a purposive sample, which means the results are not representative of all evaluations conducted in that year. As such, its use in comparing quantitative data across years is limited.
- Even though the full cohort of 40 evaluations were assessed this year, the numbers are relatively small, potentially distorting findings. The report acknowledges this where relevant, and reports numbers and percentages to further manage this limitation.
- It was not possible to gain accurate information on all aspects of the evaluations from DFAT's systems. In particular, accurate data on the cost of each evaluation and the number of person days for each evaluation was not available. This meant that the team was unable to examine the relationship between the cost of an evaluation and its quality. Rather, average costs (for 2022 evaluations) and average numbers of person days per evaluation were provided. See Annex 4 for details.
- There were time limitations on quality assurance.

3. Perceptions undermining independence.

- All team members and the Bluebird Director identified any actual or perceived conflicts of interest regarding the 40 evaluations ahead of work commencing and allocations of evaluations for assessment were made to avoid any actual or perceived conflicts of interest.
- Perceptions that the involvement of a DFAT staff person on the team might undermine the independence of the review were managed through the moderation and quality assurance processes.

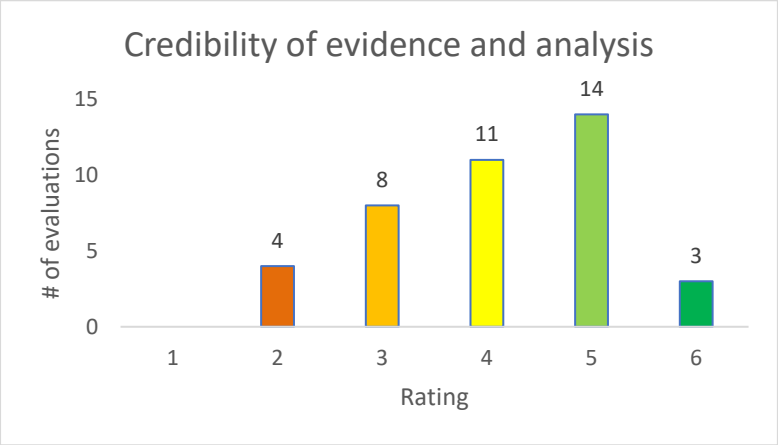
FINDINGS

Finding 1: Most evaluations (70%) provide a credible source of evidence for the development program but there is room for improvement.

The measure for overall evaluation quality is the criterion "credibility of evidence and analysis", to enable comparison between the current and previous reviews. This was used as a proxy for overall evaluation quality in the 2012 and 2014 reviews as the criterion was most strongly associated with other quality criteria in the reviews. The 2017 review identified this criterion as the best predictor of evaluation quality, given the strong positive relationship between this and the other eight criteria.[3]

The number of evaluations that were rated 1 to 6 for this criterion are shown in figure 1.

Figure 1: Overall evaluation quality 2022 evaluations



Twenty-eight (70%) of the 2022 evaluations were rated adequate or better (4 or higher) and twelve (30%) of reports were rated as less than adequate (3 or lower). Of the 28 evaluations rated adequate or better, 17 attracted a rating of good quality (5) or very high quality (6).

However, there is clear room for improvement, given 12 of the evaluations were rated less than adequate quality. Furthermore, if DFAT's strengthened approach to evaluation aspires to reaching higher quality than an 'adequate' rating, then it will also need to improve on evaluations earning a rating of 'adequate' quality (11 evaluations in 2022).

Finding 1a: There was a wide range in the quality of evaluations across the Australian development program with no discernible pattern.

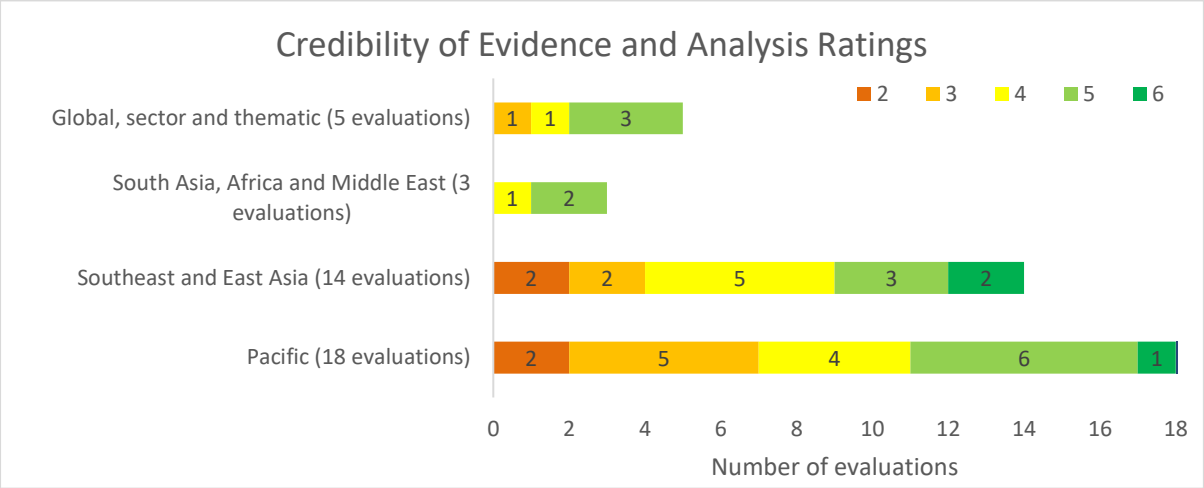
Figure 2 illustrates the wide quality range. The ratings across the two largest geographic areas – the Pacific and Southeast Asia – ranged from 2 (poor quality) to 6 (very high quality).[4] Across these regions, 61% (10) of the Pacific evaluations were rated adequate or better while 71% (11) of Southeast and East Asia evaluations were rated adequate or better.

There is no discernible pattern of good or poor quality across programs. In terms of low ratings, the four poor quality evaluations (rated 2) were from Vietnam (1), Laos (1) and Pacific Regional programs (2). The three highest rated evaluations (6, or very high quality) were from Indonesia (two of the three evaluations) and a Pacific Regional evaluation. Humanitarian evaluations also rated highly, with both rated good quality (5).

[3] Tested through correlation analysis.

[4] No evaluations were rated 1 (very poor quality) for this criterion.

Figure 2: Number of evaluations rated 2 to 6 for credibility of evidence and analysis.

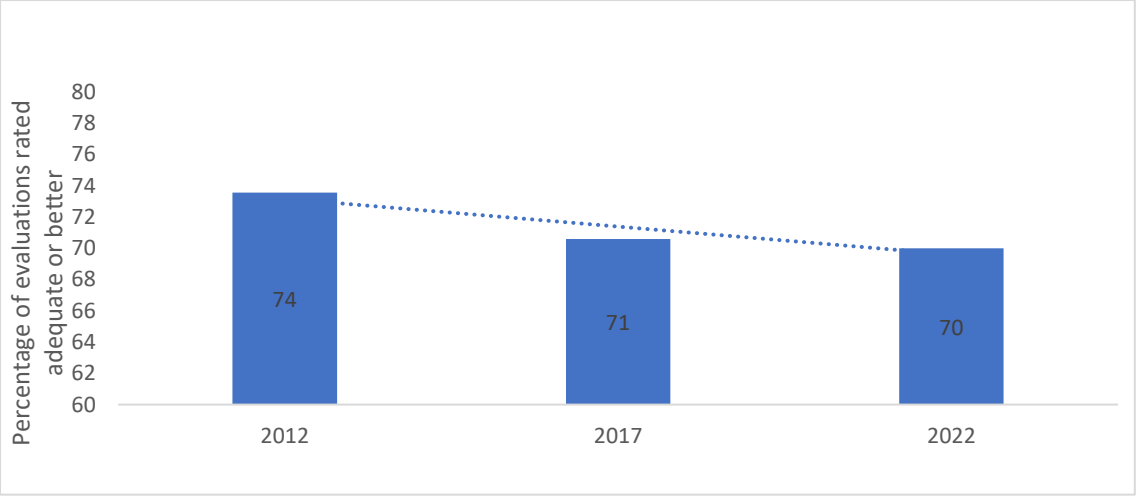


Finding 2: Compared to 2012, the quality of evaluations has declined slightly. There has been no meaningful change in overall quality since 2017.

In the last five years there has been negligible change in the overall quality of evaluations. This year 70% rated adequate or better, compared to 71% in 2017. Concurrently, there was a marginal improvement in the average quality ratings between 2017 and 2022. This is due to three of the 2022 evaluations being rated very high quality compared to none of the evaluations in 2017, and one evaluation being rated as a 1 (very poor quality) in 2017, but none in 2022.[5] This result would need to be replicated in future years to be considered a trend. On balance, the overall quality can be viewed as stable over the period 2017 to 2022 with no meaningful change.

There has however been a slight decline in evaluation quality over the past ten years as illustrated by the trend line in figure 3.[6]

Figure 3: Credibility of evidence and analysis ratings over time



[5] Annex 5 provides summary results for ratings against all criteria for evaluations in 2012, 2017 and 2022.
[6] The results for 2014 (when 77% of evaluations were found to be of adequate quality) are not included because the sample assessed in that year was not representative.

Finding 2a: Evaluations are addressing evaluation questions with increasing rigour.

This criterion considers how well the evaluation identifies appropriate evaluation questions and then answers them. It also considers whether an appropriate balance is made between operational and strategic issues.[7] The ratings for this criterion were found to correlate highly with the ratings for the credibility of evidence and analysis criterion in the 2022 reports.[8]

The ratings for the 'evaluation questions' quality criterion have improved slightly over the past ten years. In 2012, 75% were considered adequate or better and this percentage has improved to 79% in 2017 and 80% in 2022[9]. Additionally, a greater proportion of evaluations were rated good or very high quality for this criterion in 2022 (53%) compared to 2017 (32%).

Better quality evaluations:

- Had clear and appropriate evaluation questions, all of which were answered.
- Had evaluation reports that were often structured around answering the evaluation questions.
- Were clear in their judgements, with any information gaps explained.
- Identified key strategic issues while also addressing operational matters.

Poorer quality evaluations:

- Did not have well formulated questions, and/or not all were addressed. For example, some evaluations did not assess progress against outcomes when it was required to assess effectiveness, without explanation as to why.
- Often had too many evaluation questions and less coherent reports which did not distinguish between significant and less significant findings.

Improvements to evaluation quality could be made by focusing on better formulation of evaluation questions and responses to them, given the high correlation between the ratings for this criterion and the credibility of evidence and analysis.

Finding 2b: Appropriate methodologies and use of sources in evaluations have not improved in recent years.

Ratings for 'appropriateness of methodology and use of sources' declined between 2017 (76% were rated adequate or better) and 2022 (65% rated adequate or better).

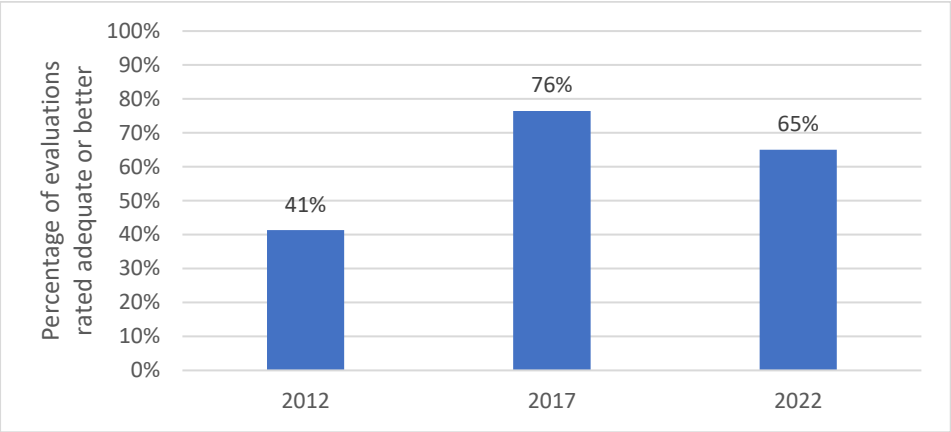
This is an important criterion, as well-chosen methods underpin the quality of evaluations. As noted in the limitations, evaluation plans were only assessed when included in the evaluation reports as an annex. Where absent, the assessment relied partly on what was included in the often brief 'methodology' section of the report. Evidence of other sub-criteria, including triangulation, appropriate sampling, means of answering each evaluation question and use of sources could be discerned to a degree from the main body of the reports.

In 2012 just 41% of evaluations had adequate methodology and use of sources (figure 4). It is possible that some of the substantial gains in quality between 2012 and 2017 may have been eroded in the last five years, but firm conclusions cannot be drawn due to the methodological challenge

[7] For more detail, see the 2023 Handbook, Annex 3, pg. 33.
[8] Established through correlation analysis.
[9] Data for 2014 (when 74% of evaluations were rated adequate or better for the 'evaluation questions' criterion) is excluded from the analysis as the sample for that year was not representative.

noted above. However, DFAT should continue to support evaluation managers to identify and assess appropriate methodologies.

Figure 4: Appropriate methodology and use of sources over time



Finding 2c: Executive summary quality has improved over time.

Better executive summaries improve the readability and potential use of evaluations. There is a clear improvement in the quality of executive summaries over time. These were rated based on how well they met the criterion: 'an accurate reflection of key findings and provides all the necessary information to enable primary users to make good quality decisions.'[10] In 2017, 73% of evaluations rated adequate or better, increasing to 83% in 2022.[11] Furthermore, more evaluations were being rated as higher quality for this criterion in 2022 than in 2017. For example, four evaluations rated very high quality in 2022 compared to none in 2017.

Finding 3: The 2022 evaluations include three of very high quality, compared to none in 2017.

This review identified several examples of good practice among the evaluations. The team recommended seven, drawn across different geographic areas and different aspects of good practice, as examples of good practice to be made available to future evaluation managers (see Annex 6).

Finding 3a: DFAT-led evaluations were slightly higher quality than partner-led or joint evaluations.

Of the 40 evaluations, 31 were DFAT-led and nine either joint or partner-led.[12] This proportion (77.5%) is comparable to the proportions in 2017 (79%) and 2012 (83%). In 2022 DFAT-led evaluations were slightly higher quality (71% adequate or better) than the combined joint and partner-led evaluations (67%). The finding that DFAT-led evaluations were stronger is similar to 2017, where 74% were rated adequate or better compared to 57% of partner-led or joint evaluations. One of the key messages from the 2017 review, that evaluations actively managed by DFAT are more likely to be good quality, remains relevant for the 2022 evaluations.

[10] 2023 Assessment Template. More detail provided in 2023 Handbook, see Annex 3.
[11] Although 2014 is excluded from the analysis, 73% of evaluations were rated adequate or better for this criterion in 2014. The quality of the executive summary was not considered in the 2012 review of evaluations.
[12] A partner-led evaluation is where DFAT relies on the evaluation process of another aid partner, such as an NGO or other donor. DFAT has no substantive input to the terms of reference, selection of the evaluation team etc. A joint evaluation is where DFAT works together with a partner (eg NGO or other donor) on the evaluation. DFAT may be the lead or an equal partner on the evaluation. A DFAT-led evaluation is where there is no involvement from another partner in the evaluation process.

Finding 3b: Very few evaluations included a DFAT staff member as a substantive member of the team, but this didn't seem to affect the quality of the evaluation.

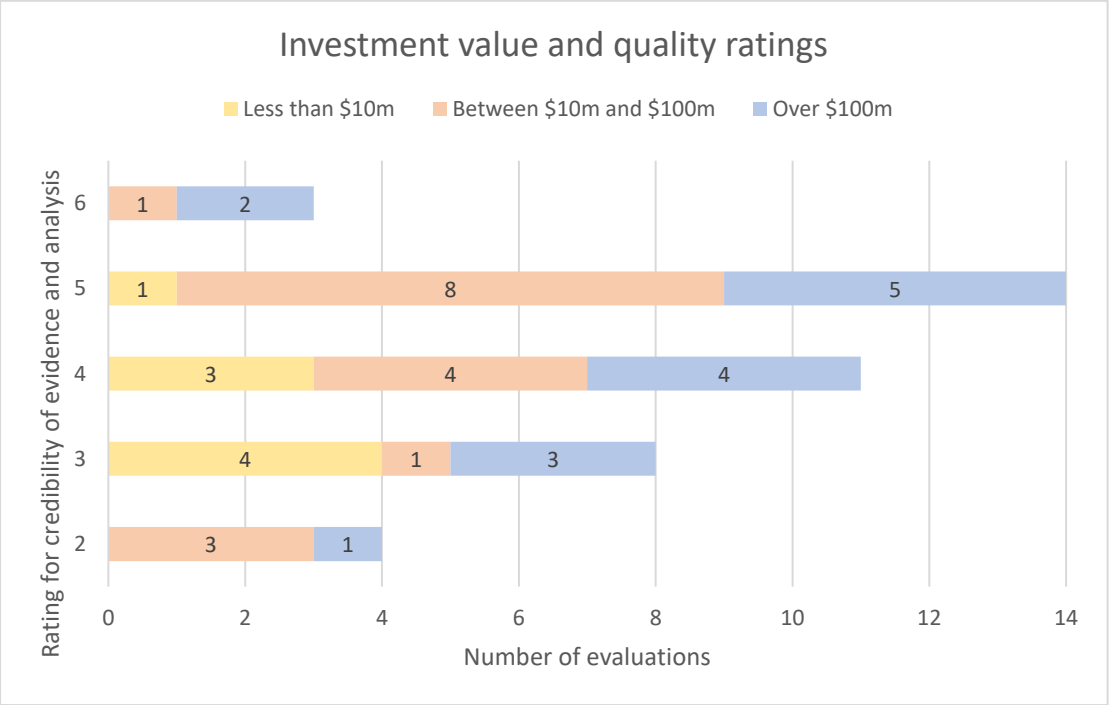
DFAT was frequently involved with evaluation oversight, such as drafting Terms of Reference, contracting, assembling teams, and reviewing evaluation outputs, but a DFAT staff member was included on the evaluation team in only three of 40 evaluations (7.5%). This is a significant change from 2017 when DFAT staff members played substantive roles as a member of the team in 11 evaluations (32%).

The review of 2017 evaluations showed that the inclusion of DFAT staff member as a member of the evaluation team corresponded with only a marginal increase in the number of adequate quality evaluations. In 2022 there was no discernible difference in the quality of the evaluations. There is no evidence from the 2017 or 2022 reviews that having DFAT staff as a substantive team member improves the quality of evaluations.

Finding 3c: Evaluations of smaller investments (less than $10m) tend to have poorer quality evaluations.

Eight of the investments evaluated in 2022 were valued at less than $10 million.[13] Half of the evaluations for these smaller-value investments (four out of eight) were rated as less than adequate (see Figure 5).

Figure 5: Evaluations of lower value investments tend to rate lower quality.



No clear conclusions can be drawn regarding the evaluation quality of the 32 larger sized investments, comprising 15 high value investments (valued at over $100 million) and 17 mid-range investments (between $10 million and $100 million). Twenty-four of the 32 (75%) had evaluations of adequate quality or better. All except one of the good or very high quality evaluations (17 altogether) were of high value and mid-range investments. However, these high value and mid-range investments also included four evaluations of poor quality (rated 2).

[13] Investment values are sourced from DFAT's aid management system (AidWorks) and sometimes cover multiple activities.

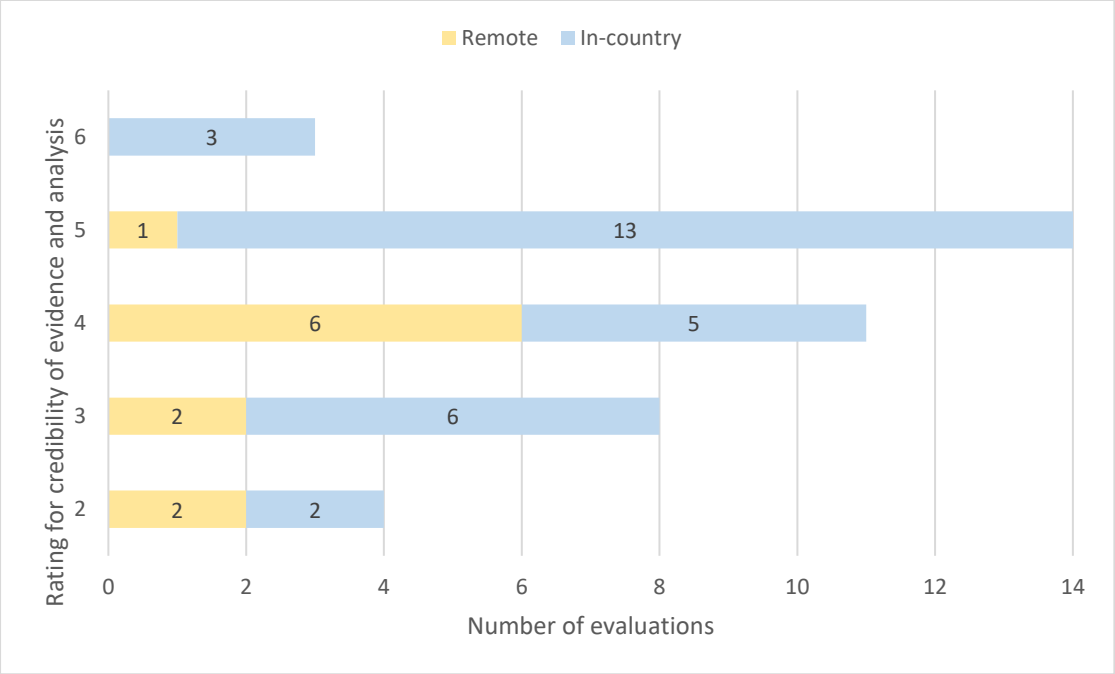
The reason why lower value investments tend to have lower quality evaluations is unclear. It could be because there are lower quality assurance requirements for investment designs under $10m and these investments are less likely to have a good M&E system. The relationship between the value of the investment and the quality of the evaluation is an area that warrants further investigation.

Finding 3d: Evaluations conducted remotely tended to be lower quality than those conducted at least partially in-country.

Of the corpus, 11 evaluations were conducted remotely with no in-country interviews or focus group discussions. These were largely due to travel restrictions in response to the COVID-19 pandemic. The remaining 29 evaluations conducted face-to-face discussions fully or partially in-country.[14]

While the numbers are small, they indicate the remotely conducted evaluations were, on average, poorer quality (see figure 6). Only one of 11 remote evaluations was assessed as good quality, compared to more than 50% (16 out of 29) in-country evaluations rated as good or very high quality.

Figure 6: Remotely conducted evaluations were lower quality than in-country evaluations.



[14] An example of an evaluation conducted partially in-country would be in a multi-country program, interviews may have been conducted in-country for one or more countries and remotely for the remainder.

DISCUSSION

This is the first *annual* review of the quality of DFAT evaluations required by Australia's International Development Policy and Delivery Framework. Until now, DFAT has undertaken these reviews periodically, but not annually. Part of the brief for this review was to consider ways in which DFAT's evaluation quality might be assessed through an efficient annual process.

Our team canvassed and considered methods to achieve this. These are presented under five main topics.

1. Adopting broader perspectives of quality and use

The quality and use of evaluations can be viewed through several overlapping perspectives[15]:

1. The quality of the evaluation *process*, such as the degree of influence the evaluation process had on the ways of thinking and perspectives of program and partner staff and whether this brought about subsequent changes in implementation practices.
2. How well the written evaluation *report* demonstrates the quality criteria for good evaluation practice.
3. The degree to which the evaluation *report* is used to *influence* current or new interventions, for example by implementing specific recommendations.
4. The *influence* of *evaluation findings* on the broader development cooperation community, including other areas of DFAT, typically through publication and dissemination of reports or syntheses of findings and learnings.

This review focused on the second point, while DFAT undertook a parallel review of use, according to the third point. It is worth considering how to assess the first point on the quality of evaluation processes, as these can often be the most influential part of an evaluation for implementing staff. The easiest way to find out the degree of influence would be to include interviews with relevant program and partner staff after the evaluation.

There are several ways to assess the broader influence of evaluations on the development cooperation community captured in point four above. First, the publication of evaluations is important. While encouraging publication of evaluations was a heavy focus of the 2012, 2014 and 2017 reviews of evaluations, it was not a focus of this 2022 review. The Aid Evaluation Policy released in 2016 mandated publication of all evaluations with a management response. The implementation of this policy has resulted in all evaluations being published, a notable achievement.

Second, making evaluations more broadly available, digestible and relevant could involve presenting findings at conferences and other forums, or publishing syntheses of evaluation findings and learnings, by sector, cross-cutting topic or geography. DFAT has conducted several such syntheses in the past, such as two learning papers on policy influence and promoting gender equality as part of the 2017 review.[16] There is opportunity for DFAT to share findings and learnings from evaluations with its own staff and to reach this broader audience. The review of quality and use could extend to an assessment of the quality and influence of this dissemination.

Taking a broader perspective on quality and use of evaluations could provide a more robust understanding of the quality and use of DFAT evaluations, and ensure efforts to improve these are directed to weaknesses and gaps that are identified. Even if all four perspectives are not addressed

[15] These perspectives are noted in evaluation theory. See for example, Weiss, C. (1998), Have We Learned Anything New About the Use of Evaluation?, American Journal of Evaluation 19(1).
[16] Policy Influence and Promoting Gender Equality.

annually, it would be useful to consciously define the scope of the annual reviews of quality and use within this framework.

2. Ways to improve review methodology

If DFAT adopts broader definitions of quality and use described above, then the review methodology would need to be revised accordingly. For example, DFAT could consider including interviews with implementing staff and partners to capture a fuller understanding of evaluation quality beyond just an assessment of the report.

A priority for the 2024 annual review is to update the methodology to align with DFAT's most recent standards. In December 2022, DFAT introduced an updated set of Design and M&E Standards. The evaluation quality criteria, assessment template and handbook need to be revised to match the standards for any future review.

The following learnings from this year's exercise should also be applied to streamline the process and further enhance consistency across assessors:

- Reduce the number of criteria to approximately 5 key areas and ensure no duplication of sub-elements within the criteria.
- Retain credibility of evidence and analysis as the core criterion.
- Develop a single overarching question for each criterion to help the assessors finalise the rating, having considered all other aspects of the criterion.
- Include assessment of TOR and evaluation plans (which could be integrated with the work of evaluation panels described above). If this is not possible, ensure all criteria are fully assessable without this information.
- Maintain a small team of approximately five people (having sufficient evaluators helps to ensure a robust moderation, but not too many to make calibration across the team members too complicated).
- Enhance quality assurance of the assessments, and include an explicit capacity building element, where required, for any evaluators on the team.

3. Using an integrated rather than a standalone process

To date, each of the evaluation reviews have been a standalone exercise where either an internal or external team has assessed evaluation reports within a fixed time frame. In future, there may be efficiencies in combining quality and use assessments with ongoing evaluation quality improvement and capacity building in DFAT throughout the year.

The 2022 cohort includes 12 evaluations rated inadequate and 11 more of 'adequate' quality rather than good or very high quality. This suggests that current quality assurance systems are not picking up issues early for all evaluations. To tackle this and enable consistent application of the Design and M&E Standards, DFAT Evaluation Managers should be given additional support for the types of evaluations that this review identified as poorer quality (those of investments under $10m, conducted remotely, or partner-led/joint evaluations).

Quality gains could be achieved by reviewing the Terms of Reference, Evaluation Plans and draft reports of evaluations early on, to allow rapid responses to any quality issues flagged in these products. Such a system could provide quality assurance and improvements in real time while collecting ongoing data on the quality of these products to feed into the annual review. This would help to triangulate with data from the evaluation reports for the annual review.

One option here is to enhance the work of ongoing quality assurance panels already contracted to program areas, to improve the consistent application of DFAT standards in assessing TOR, Evaluation Plans and Draft Reports. Where there are gaps in this system, the Development Evaluation and Assurance Section could appoint a select panel of evaluation specialists to work alongside them in assessing TOR, Evaluation Plans and Draft reports. This panel could help to strengthen the evaluation skills of both program areas and Development Evaluation and Assurance Section staff as part of their brief. Collectively these panels could become part of DFAT's evaluation strengthening approach. Meetings of the panels could, for example, promote consistency in applying DFAT's standards, support DFAT's capacity building efforts in evaluation, and provide data for the annual review.

4. Sampling considerations

If a similar number of 40 evaluations is undertaken each year within DFAT, an obvious consideration for streamlining future reviews would be to review a sample of the evaluations undertaken rather than the whole population. This was done in 2014 with a purposive sample. Such a sample, however, is not representative of the whole cohort of evaluations. For accurate results from small population sizes, such as 40 evaluations in this case, a random sample would need to include almost the full cohort of 40 evaluations. Given this, it would make more sense to include all evaluations.[17]

Alternatively, using a purposive sample based on geography or sector would allow a detailed look at a particular aspect of DFAT's work. It would, however, make year-on-year comparisons difficult, and the findings and recommendations on the quality of evaluations would likely not be universally applicable. In making the decision as to whether all evaluations should be assessed, or a sample, DFAT will need to be clear on what will most likely provide the best information to help in strengthening evaluation practice, quality and use.

5. Composition of the review team

Reviews of evaluations have been undertaken both by internal teams and external teams or a combination. The main advantages of externally contracted consultants are that experienced evaluation experts are more able to apply evaluation quality criteria consistently and provide a high-quality report within a short time frame. The disadvantages of external contracting include the expense. For the current exercise, the amount of time to review an evaluation report and complete the quality assessment template was, on average, 8.3 hours per evaluation report.[18] Notwithstanding, the use of internal DFAT staff also comes at a cost as it is a large time commitment.

The contracted arrangement for this review included four external consultants and one staff person from the Evaluation Section in DFAT. A strength of this arrangement was that it provided an opportunity to develop the skills of the DFAT staff member in recognising and supporting good quality evaluations. That said, as this exercise is about applying deep evaluation experience, there are risks in including less experienced people on the team, which were mitigated to a degree by providing additional support for the team member in this review. This arrangement also had the advantage of the DFAT staff member undertaking the reviews of the two evaluations in the pool which were exempted from publication and would not have been able to be included in the review otherwise.

[17] 'Most statisticians agree that the minimum sample size to get any kind of meaningful result is 100. If your population is less than 100 then you really need to survey all of them.' tools4dev.org

[18] Based on timesheets for 4 consultants and 28 evaluation reports. Note the range was from 5 hours per assessment to 14 hours.

RECOMMENDATIONS

To address the findings of this review and options for future reviews, it is recommended that the Development Risk, Implementation and Evaluation Branch in DFAT:

1. **Continue to support DFAT program areas to manage high quality evaluations, with a particular focus on:**
    a. Meeting DFAT's Design and M&E Standards such as by formulating relevant evaluation questions, having an appropriate methodology to answer them, ensuring the evaluation report clearly addresses each question and provides sufficient evidence to support its findings.
    b. Sharing good practice examples.
    c. Promoting and building DFAT staff capacity to actively manage evaluations.

2. **Coordinate early review of evaluation TORs, evaluation plans and draft reports by expert providers to improve evaluation quality in a timely manner, and integrate these assessments with the annual review.**
    a. This could be done by enhancing the work of existing quality assurance panels already contracted to program areas and supporting DFAT's Development Evaluation and Assurance Section with additional expert providers as needed.
    b. These expert providers could promote consistency of assessments, support DFAT's capacity building efforts, and provide data for the annual review.

3. **Articulate a broader perspective on quality and use of evaluations and define the scope of each annual review within this framework.**
    a. Adapt the methodology for the annual review of the quality of DFAT evaluations to match the revised scope, for example include interviews with staff and partners to capture a fuller understanding of evaluation quality beyond the report assessment.
    b. Further update and streamline the methodology by revising the evaluation criteria, assessment template and handbook to meet the latest DFAT Design and M&E Standards and to address some of the methodological limitations of this review and learnings from implementation. For example, reduce the number of criteria to approximately 5 key areas and ensure the sub-elements within the criteria are not duplicated.

ANNEX 1: REVIEW PLAN

Quality Review of DFAT 2022 Development Evaluations Review Plan

1. Introduction

DFAT has contracted Bluebird to conduct the Quality Review of Development Evaluations completed in 2022 (the review). This will examine the quality of development program evaluations completed in 2022. The review follows on from the previous three reviews of Program Evaluations, which examined program evaluations completed in 2012, 2014 and 2017.

2. Background

The Australian Government will release a new international development policy in mid-2023. It will commit to a strengthened approach to evaluation, underpinned by a focus on quality and use. An annual review of quality and use of evaluations is expected to contribute to measuring the performance and delivery of the new development policy.

DFAT conducts around 40-50 program evaluations each year, in line with the Development Evaluation Policy ('the policy') introduced in 2016 and updated in 2020. The Policy makes clear that "use is the driving force behind our evaluations".

The Policy focuses on three areas to ensure evaluations are useful: prioritisation, quality and systems which facilitate use. The former Office of Development Effectiveness (ODE) completed three reviews which assessed the quality of program evaluations in 2012, 2014 and 2017. These reviews applied the same criteria and methodology to enable comparison across years.

The 2017 review focussed on assessing the impact of the revised Aid Evaluation Policy (2016) on evaluation practice, quality and use. Since the last review in 2017 there has been a significant shift in the evaluation function in DFAT with the disbandment of ODE and the Independent Evaluation Committee in 2020, an increased focus on evaluation across the Australian Public Service with the Commonwealth Evaluation Policy released in 2021, and the imminent release of a new international development policy.

3. Purpose

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

The review will provide an opportunity to assess how well DFAT is implementing one of the key elements of the Development Evaluation Policy - quality - and enable a comparison of results from past reviews. Use of evaluations will be separately assessed by DFAT.

4. Objectives

The review has three objectives:

1. To better understand the practices related to and quality of independent program evaluations, and how these have changed over time by comparing to findings to those of similar reviews conducted in 2012, 2014 and 2017.

2. To inform approaches to strengthen evaluation across the department.
3. To contribute to evidence that Australia's development cooperation is informed by monitoring, evaluation and learning.

5. Scope

The review will examine all 40 independent program evaluations completed in 2022. This includes evaluations commissioned by DFAT as well as joint or partner-led evaluations included on DFAT's 2022 Annual Development Evaluation Plan. The 2022 evaluations will give the most up to date data on current evaluation practice in the department. Reviewing all 40 evaluations will allow firm conclusions to be drawn that is based on analysis of the full population of relevant evaluations.

6. Audience

The primary audiences for this review are staff from the Development Evaluation and Assurance Section, senior managers in the Development Effectiveness and Enabling Division, DFAT's Executive and the Development Policy Sub-Committee. The findings from the review will be used by these groups to inform the department's approaches, policies, guidance, training and practices in monitoring, evaluation and learning.

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

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

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 considered by the Development Policy Sub-Committee
- The full review report, including executive summary, will be published on the DFAT website.
- PRD staff and the consultant will present the review process and findings at appropriate DFAT and other forums as opportunities arise. e.g., Australasian Evaluation Society's annual conference, Annual Australasian Aid Conference, EvalNet meetings and the Australian Public Service Evaluation Community of Practice.

7. Review Questions

The review will answer the following evaluation questions.

1. What is the quality of program evaluations? Has this changed over time? How?
2. How does quality differ between different evaluation types, including partner-led, and evaluations conducted remotely or in a hybrid format?
3. To what degree do program evaluations provide a credible source of evidence for the effectiveness of Australia's development cooperation program?
4. Based on the findings of this review, what are the implications for DFAT's Development Evaluation Policy and processes?
5. Which evaluations (Terms of Reference, Evaluation Plan or Report) can be promoted as good practice examples?

8. Methodology

The review methodology is similar to that used for the 2017 review, to help enable comparison over time. However, DFAT is seeking a more streamlined process, which means there will be less data collection and analysis than in previous reviews. The process will comprise a desk review of all program evaluations completed in 2022 (numbering 40 evaluations).

The following table summarises how each evaluation question will be answered. More detail of data collection and analysis methods follow the table.

Table 1: Summary of methods to answer evaluation questions.

Evaluation question	Data collection methods	Data analysis methods
1. What is the quality of program evaluations? Has this changed over time? How?	Basic characteristics of 2022 evaluations (and the investments they relate to) collected from evaluation reports and recorded in assessment template. 2022 evaluations rated against quality criteria in assessment template by expert evaluators. Evaluation characteristics and quality summarised from the 2012, 2014 and 2017 Review of Program Evaluations	Analyse each of the criteria to establish areas where evaluation quality is high and low. Comparative tables/ figures of evaluation characteristics and quality, comparing with previous three reviews of Program Evaluations, in particular around the "credibility of evidence and analysis" criteria as a proxy of overall evaluation quality.
2. How does quality differ between different evaluation types, including partner-led, and evaluations conducted remotely or in a hybrid format?	Data on such factors and evaluation quality collected under Q1 above.	Correlation analysis to examine relationship between evaluation quality and possible factors contributing to evaluation quality. Limited narrative analysis of completed reviews and team meeting conversations.
3. To what degree do program evaluations provide a credible source of evidence for the effectiveness of the Australian aid program?	Data on evaluation quality collected under Q1 above.	Analysis of assessments against quality criterion 8, 'credibility of evidence and analysis'.
4. Based on the findings of this review, what are the implications for DFAT's Development Evaluation Policy and processes?	Insights from team members on (i). factors likely to be contributing or detracting from evaluation quality (whether considered in this year's review or not) and (ii) ways in which evaluation quality might be assessed in efficient annual processes will be discussed at selected team meetings.	Data and analysis from Q1, Q2 and Q3 and Q4 will be collated and presented back to the team in a workshop to identify or verify the implications for DFAT's evaluations policies and practices, and in particular, ways in which DFAT's evaluation quality might be assessed in efficient annual processes.
5. Which evaluations (Terms of Reference, Evaluation Plan or Report) can be promoted as good practice examples?	The assessment template includes a section where specific aspects of the evaluation can be identified as good practice. In addition, the top rating evaluations will also be considered.	A limited number of good practice examples with basic description will be proposed to DFAT based on the expert review. DFAT may then undertake further description of the stronger features of these nominated evaluations to enable their dissemination as good practice examples.

Q1: What is the quality of program evaluations? Has this changed over time? How?

Evaluation types and characteristics

As part of the streamlined approach, information collected on the types and characteristics of the evaluations will be limited to the following, including three new characteristics as indicated:

- Geographic region of evaluation
- DFAT or partner led.
- Evaluation conducted by a specialist quality assurance group eg the Quality and Technical Assurance Group (QTAG) for PNG (new)
- Remote, in-country or hybrid (new)
- DFAT staff on evaluation team, and if so, role of DFAT staff on evaluation team
- Progress or completion report
- Cost of evaluation
- Total investment value
- Investment duration (new)
- Number of evaluation questions

The question about whether the evaluation was conducted remotely, in-country, or a combination of the two (hybrid) is a new question introduced this year, as the COVID-19 pandemic had a significant impact on the way in which evaluations were conducted and DFAT would like to know if this had consequences for evaluation quality.

Another new question will be to ascertain whether the evaluation was conducted by a specialist quality assurance group, such as the Quality and Technical Assurance Group (QTAG) for PNG. The reason for this question is that there has been an increase in the number of these specialist groups since the review of 2017 evaluations and DFAT are interested to know if there is a link to evaluation quality.

Information previously collected on whether the evaluation was single or a cluster evaluation (of multiple activities), skills of the evaluation team, number of people on the team and the number of person days the evaluation took will not be collected. It is difficult to find some of this information and the value it brings to the review exercise is questionable. In the review of 2017 evaluations for example, factors of team composition, team size and duration of the evaluation could not be linked to evaluation quality.

Information about the sector of the evaluated programs will not be specifically collected. This will avoid any confusion with respect to DFAT's new development policy (not released at the time of writing), which may include new sector categories or naming conventions.

Information about whether the evaluation assessed DFAT's performance (quality) criteria, and the ratings assigned will no longer be collected as it is unnecessary for this particular exercise.

Quality

The quality of each evaluation will be assessed by the experienced evaluators using the same nine quality criteria used in all the previous reviews, which are based on DFAT's Monitoring and Evaluation Standards (prior to the updating of these standards in December 2022). The nine quality criteria are: quality of executive summary; purpose of evaluation; scope of evaluation; appropriateness of the methodology and use of sources; methods; adequacy and use of M&E;

context of the initiative; evaluation questions; credibility of evidence and analysis; and quality of recommendations.

This assessment will be guided by a manual which includes greater detail on each of the nine criteria, derived from DFAT's Monitoring and Evaluation Standards. The team members (expert evaluators) will arrive at a rating of 1 to 6 (see Table 2) and record a narrative comment in their rationale for assigning that rating, in the assessment template.

Table 2: Ratings

Satisfactory		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; does not fail in any major area	1	**Very poor quality:** does not satisfy criteria in any major area

Moderation process

These ratings will be moderated through a combination of means, to ensure the ratings across team members and across quality reviews of evaluation reports are as consistent as possible. First the team leader was a member of the team from the review of 2017 evaluations, and this provides a level of continuity that is helpful in setting the bar for expectations of quality. Second, a moderation exercise at the commencement of work (described in detail below), which involves all team members rating the same evaluation, 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 and establish the basis for quality assessments. Third, the ongoing weekly team meetings provide the means by which team members can discuss and agree on ratings which are difficult to assess, reinforcing and bringing more nuanced understandings of the interpretations and relative priorities of detailed aspects of the nine criteria.

The process for the moderation exercise will be as follows:

i. The team leader will select one evaluation for the exercise. Each team member will be provided with the package of assessment materials (comprising this review plan, the assessment template, the manual, and a copy of the applicable DFAT Design, Monitoring and Evaluation Standards). In the first instance, team members will be requested to assess the selected evaluation only against the quality criterion "Credibility of evidence and analysis".
ii. At the first team meeting, if the team members have given the sample evaluation different ratings for "Credibility of evidence and analysis", they will be invited to share and discuss the reasons why. Moderated by the team leader, this discussion will lead to an agreement on the rating to be given and the reasons why. The team will then be asked to consider the other criteria and there will be opportunity to clarify any queries regarding the template and manual. The team will be asked to complete the assessment of the sample evaluation following the first meeting.
iii. Prior to the second team meeting, the team leader will analyse the ratings and reasons from each of the team members for the remaining eight criteria and provide feedback at the second team meeting that will help pinpoint the reasons for any discrepancies in the ratings

given by different team members. In a similar process to the first team meeting, agreement on the ratings for each of the criteria and the reasons why will be reached through discussion.

Following the moderation exercise, team members will undertake the quality reviews of all assigned evaluations. The assessment templates will be shared with and collated by the team leader for subsequent analysis.

Analysis

The team leader will analyse each of the nine criteria to establish areas where evaluation quality is high and low, based on the ratings. The team leader will prepare comparative tables and figures of evaluation characteristics and quality, comparing with the previous three Reviews of Program Evaluations., in particular around the "credibility of evidence and analysis" criteria.

The measure for overall evaluation quality will be the criterion "credibility of evidence and analysis", to enable comparison between the current and previous reviews. This was used as a proxy for overall evaluation quality in the 2012 and 2014 reviews as the criterion was most strongly associated with other quality criteria in the reviews. The 2017 review undertook correlation analysis to test if this is the best predictor of evaluation quality and found there was a strong positive relationship between this and the other eight criteria.

Q2: How does quality differ between different evaluation types, including partner-led, and evaluations conducted remotely or in a hybrid format?

Using the data collected above, the team leader will undertake correlation analysis to examine any relationship between evaluation quality and different evaluation types, including partner-led, and evaluations conducted remotely or in a hybrid format. Further analysis will look for other possible factors contributing to evaluation quality.

This will be complemented by a limited narrative analysis of completed reviews (the reasons for assigning a particular quality rating) and team meeting conversations, which will seek additional insight from team members on the details of better or poorer quality in the evaluations as well as insights as to possible factors contributing to evaluation quality. Findings will be tested and validated during a team workshop to consider the draft findings.

Q3: To what degree do program evaluations provide a credible source of evidence for the effectiveness of Australia's development cooperation program?

As described above, analysis of assessments against quality criterion 8, "credibility of evidence and analysis" will provide information for the overall quality of DFAT's 2022 evaluations. This information will be compared to the findings from the previous three quality reviews.

Q4: Based on the findings of this review, what are the implications for DFAT's Development Evaluation Policy and processes?

Insights from team members on (i) factors likely to be contributing or detracting from evaluation quality (whether considered in this year's review or not) and (ii) ways in which evaluation quality might be assessed in efficient annual processes will be discussed at selected team meetings.

Data and analysis from Q1, Q2, Q3 and Q4 will be collated and presented back to the team in a workshop to identify or verify the implications for DFAT's evaluations policies and practices. These will be presented in the report as a limited number of recommendations.

Q5: Which evaluations (Terms of Reference, Evaluation Plan or Report) can be promoted as good practice examples?

The assessment template includes a section where specific aspects of the evaluation can be identified as good practice. These will be reviewed by the team leader. In addition, the top rating evaluations will also be considered by the team leader. A limited number of good practice examples with basic description will be proposed to DFAT based on this process. DFAT may then undertake further description of the stronger features of these nominated evaluations to enable their dissemination as good practice examples.

9. Team roles

The assessments of each evaluation report will be undertaken by an expert evaluator with considerable international development and DFAT experience. The team comprises four team members (three external and one DFAT assessor from the EVS Team) and a team leader. The roles and responsibilities are as follows.

Team Leader

- Oversee review process, lead the team, troubleshoot.
- Moderate consistency of assessments across the team
- Conduct data analysis (quantitative and qualitative), prepare tables, graphs and figures.
- Draft the Review Plan and methodology for the review, update the template, manual and draft and finalise the final report.
- Liaise with DFAT

Team Members (expert evaluators)

- Assess the quality of program evaluations (up to 11 evaluations per team member) by completing the assessment template for each assigned evaluation.
- Participate in moderation and weekly meetings and final workshop to share insights and help validate findings and propose recommendations.

10. Limitations and risks

i. Consistency of assessments across the team and across years

The quality of program evaluation reports will be assessed by four team members using the assessment template at Annex A. To ensure the findings of the review are credible, it will be important to ensure team members assess program evaluations relatively consistently, and that this assessment can be reasonably compared to assessments of 2017, 2014 and 2012 evaluations. Consistency across years is a risk, but the use of the same nine criteria and DFAT Design, Monitoring and Evaluation Standards reduces this risk. Also, the team leader's involvement in the 2012 and 2017 reviews 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.

ii. Perceptions undermining independence

All team members will be requested to identify any actual or perceived conflicts of interest regarding the 40 evaluations ahead of work commencing. The team leader will document these conflicts or potential conflicts and ensure that team members are not allocated the relevant evaluations to assess. Team members should raise a perceived conflict of interest at any time it becomes apparent. Note that the two unpublished evaluations will be assigned to the DFAT staff member of the team.

The involvement of a DFAT staff person on the team might be perceived to undermine the independence of the quality review. This will be managed through the moderation processes described above. Also, the majority of the team, including the team leader, are external to DFAT and this will mitigate this risk.

11. Schedule

The review will take place from July to October 2023, with the absolute deadline for the final report being 31 October 2023. The table below indicates the main tasks, persons responsible and approximate timeframe for the review tasks.

Table 2: Review Tasks, Responsibilities and Timeline

Task	**Person(s) responsible**	**Approximate dates for completion**
Obtain declarations of conflict of interest	Team Leader	13 July
Inception meeting to agree review plan, templates, handbook, and final report format	Team Leader and DFAT	13 July
Prepare first team meeting: send out package of information to team	Team Leader	20 or 21 July
First Team Meeting	Team Leader and Review Team	15 August
Main Moderation meeting	Team Leader and Review Team	22 August
Assess the quality of each of remaining 37 program evaluations and record in template	Review Team members	23 August to 29 September
Team Meetings to moderate rating consistency, trouble shoot, share insights	Team Leader and Review Team	29 Aug, 5 Sep, 12 Sep, 19 Sep
Analysis and Draft preliminary findings	Team Leader	2 Oct
Team workshop, including DFAT, to discuss analysis and agree on key findings	All team members plus DFAT	5 Oct
Draft report	Team Leader	13-Oct-23
DFAT review of report	DFAT	20-Oct-23
Finalise report	Team Leader	31 October 2023

12. Outputs

Outputs will include:

- This review Plan outlining the detailed methods to be used for the review, including an Assessment Template and Handbook.
- 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)
    - A DFAT-provided Annex on the use of evaluations (or similar reference)
- Detailed records of data collected.

ANNEX 2: PROGRAM EVALUATIONS COMPLETED IN 2022

PACIFIC (18 evaluations)

Country/Region	Evaluation
Papua New Guinea	Final Review of PNG-Australia Governance Partnership
Papua New Guinea	Education Emergency Response and Recovery Plan Independent Evaluation
Papua New Guinea	Independent Review of South Fly Resilience Plan
Papua New Guinea	Review of the Markets, Economic Recovery, and Inclusion Program (Phase One)
Papua New Guinea	Review of the PNG-Australia Transport Sector Support Program Phase 2 (TSSP2)
Papua New Guinea	Australia, New Zealand, International Finance Corporation: Papua New Guinea Partnership Midterm Evaluation
Papua New Guinea	Justice Services and Stability for Development Program (JSS4D) Mid-Term Review
Pacific Regional	Australia-SPC Partnership Evaluation
Pacific Regional	Australian Infrastructure Financing Facility for the Pacific Two-Year System-Wide Review
Pacific Regional	Australia's COVID-19 Response Package for the Pacific and Timor-Leste Independent Review 2020-2022
Pacific Regional	Mid-Term Review Report of the Pacific Insurance and Climate Adaptation Programme
Pacific Regional	End of Investment Evaluation: Pacific IUU Fishing
Pacific Regional	Pacific Digital Economy Programme Mid-Term Review
Pacific Regional	*Pacific Connect evaluation (exempt from publication)*
Solomon Islands	Ombudsman Twinning Support Independent Review
Solomon Islands	Review and Evaluation of the Performance of Sustainable Transport Infrastructure Improvement Program (STIIP) and the National Transport Fund (NTF) in the Solomon Islands
Fiji	Strategic Review of the Fiji Health Program
Nauru	*Every Life Matters:* Review of DFAT Health Investments to Nauru

SOUTH EAST AND EAST ASIA (14 evaluations)

Country/Region	Evaluation
Indonesia	Independent Strategic Review of Innovation for Indonesia's School Children Phase 2 and Rural and Remote Education Initiative for Papua Provinces Phase 3
Indonesia	Penyediaan Air Minum dan Sanitasi Berbasis Masyarakat (PAMSIMAS) Final Independent Evaluation
Indonesia	Australia-World Bank Indonesia Partnership Independent Mid-Term Review
Timor-Leste	Joint Independent Evaluation - Timor-Leste Police Development Program
Timor-Leste	Partnership for Human Development Mid-Term Review
Vietnam	Aus4Reform Review
Vietnam	Aus4Innovation Mid-term Review
Cambodia	Australia-Cambodia Cooperation for Equitable Sustainable Services (ACCESS) End of program evaluation
Cambodia	Ponlok Chomnes Independent Strategic Review
Laos	BEQUAL Phase 1 Independent End of Program Review
Mongolia	Australia Mongolia Extractives Program (AMEP) II Mid - Term Review
ASEAN and Mekong	ASEAN-Australia Digital Trade Standards Initiative Mid-Term Review
ASEAN and Mekong	Mid-Term Review: ASEAN Australia Smart Cities Trust Fund (AASCTF)
ASEAN and Mekong	ASEAN-Australia Counter-Trafficking Program Mid-Term Review (MTR)

SOUTH ASIA, AFRICA AND THE MIDDLE EAST (3 evaluations)

Country/Region	Evaluation
Bangladesh	Program Completion Review of the Strategic Partnership Arrangement (SPA) Phase 2 in Bangladesh between DFAT, FCDO and BRAC
Sri Lanka	Independent Evaluation of Women in Work (WIW) Program, Sri Lanka
Afghanistan	*Ending Violence Against Women evaluation (exempt from publication)*

GLOBAL, SECTOR AND THEMATIC (5 evaluations)

Program	Evaluation
Humanitarian	Australia Assists End of Program Evaluation
Humanitarian	Review of the Humanitarian Logistics Capability
Australia-NGO Cooperation Program	Independent Evaluation of the Australian NGO Cooperation Program (ANCP)
Health	Evaluation and Forward Scoping for the Therapeutic Goods Administration's Regulatory Strengthening Program and the Australian Expert Technical Assistance Program- Regulatory Support and Safety Monitoring
Education	Mid-term evaluation of the Global Education Monitoring (GEM) Centre Phase 3

ANNEX 3: HANDBOOK

Review of Program Evaluations completed in 2022

Handbook for completing the Assessment Template

The Excel Assessment Template is in two parts – a cover sheet and quality criteria. This handbook provides additional detail to reviewers to aid consistent completion of the assessment. For the quality criteria, reference is also be made to the DFAT Monitoring and Evaluation Standards, April 2017.

Cover sheet

Reviewer	Insert your name as the reviewer of quality of this evaluation.
Investment name	In the first column write the name of the investment being evaluated. If more than one investment, list the investment names.
Investment value	Insert value of investment in the first column if it is stated in the report. If more than one investment is evaluated, list the values of each investment.
Investment Dates	In the first column, record the start date of the investment. If multiple investments, list the start dates. In the Notes column, record the end date of the investment. If multiple investments, list the end dates.
Evaluation purpose	In the first column, record: • P if the evaluation is a progress report • C if the evaluation is a completion report
Evaluation is partner-led (P), joint (J) or DFAT-led (D)?	In the first column, record P, J or D to reflect whether the evaluation was partner-led, joint, or DFAT led. A partner-led evaluation is where DFAT relies on the evaluation process of another aid partner, such as an NGO or other donor. DFAT has no substantive input to the terms of reference, selection of the evaluation team etc. A joint evaluation is where DFAT works together with a partner (eg NGO or other donor) on the evaluation. DFAT may be the lead or an equal partner on the evaluation.

	A DFAT-led evaluation is where there is no involvement from another partner in the evaluation process. If you record 'P' or 'J', in the Notes column outline the partner(s) involved in the evaluation.
Evaluation cost (if available)	DFAT to supply, leave blank.
Evaluation team: DFAT staff member included?	In the first column, record: • 'Y' if a DFAT staff member is included in the evaluation team • 'N' if a DFAT staff member is not included in the evaluation If you record 'Y', in the Notes column outline the main role of the DFAT staff member, eg an observer, an active team member with a substantive role in data collection, report writing etc.. If the role of the DFAT staff member is unclear, record "unclear" in the Notes column.
Was the evaluation conducted by a specialist DFAT quality assurance group?	In the first column, record Y if the evaluation was conducted by a specialist DFAT quality assurance group, for example the Quality and Technical Assurance Group (QTAG) for PNG. If not, enter N. If Y, in the Notes column enter the name of the group.
Was the evaluation conducted remotely (R), in-country (C) or a hybrid of the two (H)?	In the first column, record R if the evaluation was conducted remotely, C if it was conducted in-country and H if the evaluation was conducted partly in country and partly remotely. If R or H, in the Notes column, enter details about why it was conducted this way and the extent this affected the evaluation, if mentioned.
Number of evaluation questions	In the first column, record the number of evaluation questions. Sub-questions should be counted as evaluation questions. In the second column additional useful information could be recorded, for example how many are main questions and how many are sub-questions.

Quality criteria

Each box below describes what reviewers should look for when assessing evaluations against the quality criteria.

For each criterion, a rating between 1-6 should be given according to the rating scale below.

The detailed descriptions of the nine criteria generally outline what a good quality evaluation looks like. The rating scale below outlines how a criteria is rated as either satisfactory or less than satisfactory.

Ratings

Satisfactory		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; does not fail in any major area	1	**Very poor quality:** does not satisfy criteria in any major area

N/A: The criterion does not apply to the evaluation

1) Purpose of evaluation The purpose of the evaluation is provided, including the overall purpose and primary users of the information
The evaluation products clearly identify the overall purpose(s) and objective(s). It shows which purposes are of most importance – eg accountability, investment improvement, knowledge generation/learning. The *primary users* of the information are identified. They are identified by title not only organization. For example, "DFAT" is made up of senior executive, desk officers, senior managers and initiative managers. "The Contractor" is made up of head office personnel, implementation managers and advisers. It is clearly articulated that the report will be published on the DFAT website and there are clear instructions on how sensitive information is to be communicated. Evaluation products describe any previous evaluations of the investment, including a summary of findings and if recommendations have been implemented. Evaluation products also describe the relationship between the previous and current evaluations. This relationship appears reasonable and the different evaluations complement each other. For example, an early evaluation may focus on implementation/program management while a later evaluation may focus on whether outcomes have been achieved. Source: M&E Standards 4.2, 5.2 and 5.3

2) Scope of evaluation

The scope and questions matches the evaluation time and resources; methods are defined and roles of the team, DFAT management and others are set out. This criteria relates to the planning of the evaluation more than the execution. The TORs (largely) and the evaluation plan (if provided in the Annexes) should be the main reference point/s for assessing this criteria. The evaluation resources, time, methods and skills/roles of the team should match the purpose and questions of the evaluation.

Scope and timing

The scope of the questions is suitable for the time and resources available for the evaluation. The scope aligns to the purpose of the evaluation. There are sufficient number of days allocated to answer all the evaluation questions, as well as to work together as a team to process and discuss findings.

Time has been allocated to *reviewing* investment documentation (approx. 2 days) as well as time to *appraise* any key documents such as gender equality, disability and social inclusion, or sustainability strategies, or the M&E system (often a day per document for full appraisal).

The number of days allocated to completing the evaluation report reflects: a) the scope of the evaluation questions; b) the complexity of the issues that have emerged; c) the number of people contributing to the writing of the report; d) team reviewing and discussions of the final draft.

Some broad timing guidelines are:

- Typically, a 12-day in-country evaluation can only address four or five broad questions.
- Most 60 minute interviews with a respondent cover no more than four or five key topics; less if translation is required.

Methods

Evaluation products (particularly the evaluation plan) show how each of the evaluation questions will be answered by describing the methods that will be used to collect the information.

Consideration is given to the design of data collection methods that are responsive to the needs, rights and security of respondents, with special consideration given to the needs of any special sub-groups (eg women, people with disabilities).

The design of major evaluation activities/studies are annexed and include tools such as interview guides or questionnaires.

Summary statements of methods that are not linked with specific evaluation questions are not considered adequate.

Team roles

Evaluation products (particularly the ToRs and evaluation plan) outline how each team member will contribute and their responsibilities. This may include responsibility for particular evaluation questions and for writing particular parts of the report.

Source: M&E Standards 4.7, 4,10, 4.13, 4.16, 4.17, 4.18, 4.19, 5.9, 5.17, 5.18

3) Appropriateness of the methodology and use of sources

The methodology includes justification of the design of the evaluation and the techniques for data collection and analysis. Methods are linked to and appropriate for each evaluation question. Triangulation is sufficient. The sampling strategy is appropriate (where applicable).

Methodology should be appropriate and proportionate to the value, complexity and context of the investment, and purpose and scope (including evaluation questions) of the evaluation.

Limitations to the methodology and any constraints encountered are described.

Ethical issues such as privacy, anonymity and cultural appropriateness are described and addressed.

The main reference points for assessment of this criteria will be the methodology section of the report and the evaluation plan (if attached as an Annex). If the stated methodology was not able to be used, including evaluation questions not addressed, then this should be explained.

Methods

Justification is provided by the data collection and analysis techniques chosen. The methods described can reasonably answer the evaluation questions posed. For example, a focus group discussion would be most unlikely to answer a sensitive question.

Evaluation products (particularly the evaluation plan) describe how data will be analysed. Consideration is given to the analysis of disaggregated data for gender and other relevant sub-groups where possible.

Triangulation (the use of a range of methods and/or sources of information to come to a conclusion or result) is proposed. In a typical DFAT evaluation, this might include discussion of similar questions across a range of different respondents within and across different organizations or target beneficiary groups (particularly special sub-groups), or use a number of methods to examine the same issue. It is not sufficient to state that triangulation will be used if this is not demonstrated in the evaluation design.

Appropriate sampling strategies are chosen and justified. For short reviews that rely on analytical rather than statistical inference, purposeful sampling will be appropriate and could include maximum variation, a critical case, or a typical case. Efforts should be made to avoid relying on a convenience sample which is likely to be unrepresentative of the population of interest.

Limitations

Key limitations are summarised in the evaluation report to enable the reader to make appropriate decisions. Where necessary the author has provided specific guidance of where the reader ought to be cautious about the findings.

Ethical issues

Ethical issues and how they will be addressed are identified. For most of the evaluations and reviews conducted by DFAT, this will mostly be around privacy and confidentiality issues. The plan identifies how these will be addressed when data are collected, stored and reported. In particular, assurances about anonymity must be honoured and data stored and reported in ways that do not inadvertently identify informants.

Sources: M&E Standards 5.10, 5.11, 5.12, 5.13, 5.14, 6.3

4) Adequacy and use of M&E

The adequacy of M&E data/systems are described. The evaluation makes use of the existing M&E data.

Evaluation products provide a broad description of what data is available from the investment's M&E system. How this data will be used during the evaluation is discussed.

If existing data from the investment's M&E system won't be used, a brief explanation as to why is provided.

The use of data from the investment's M&E systems appears reasonable, based on the quantity and quality of data available. For example:

- we would expect good quality data from a good M&E system would be used for the evaluation
- we would expect poor quality data from a sub-standard M&E system would not be used for the evaluation.

5) The context of the initiative

The context of the initiative is described (including policy, development and institutional context) and its influence on performance is assessed.

Evaluation products identify relevant aspects of the context within which investments are implemented. These might include geographic, cultural, gender, political, economic or social context.

Sufficient information is presented to allow the reader to understand the relationship between the initiative and its context.

The report addresses: a) how the context may have affected the achievement of outcomes (both supportive and inhibiting); and b) the extent to which the investment may have had any effect on the context.

Important emergent risks are identified.

Source: M&E Standard 6.11

6) Evaluation questions

The report identifies appropriate evaluation questions and then answers them. An appropriate balance is made between operational and strategic issues.

The evaluation report clearly addresses all questions from the ToRs/Evaluation Plan. The report does not need to be a mechanical presentation of the evaluation questions, but it should be relatively easy to negotiate the report and find relevant information about specific questions. Where there are gaps, these have been explained. DFAT's information needs, as set out in the Terms of Reference and Evaluation Plan, have been met.

The report addresses the full range of issues identified in response to the TOR and other critical issues that have emerged. Strategic direction or other higher order issues related to the investment have been given adequate space, and minor technical issues are treated in a more limited fashion.

Source: M&E Standards 6.5, 6.7

7) Credibility of evidence and analysis

Findings flow logically from the data, showing a clear line of evidence. Gaps and limitations in the data are clearly explained. Any assumptions are made explicit.

Conclusions, recommendations and lessons are substantiated by findings and analysis. The relative importance of findings is stated clearly. The overall position of the author is unambiguous

In assessing outcomes and impacts, attribution and/or contribution to results are explained. Alternative views / factors are explored to explain the observed results.

Major criteria:

- The presentation of evidence is credible and convincing. Key findings are clearly substantiated by evidence and the sources of data are provided. Gaps/limitations in the data are explained.
- Evidence has been coherently considered from a range of sources, including key stakeholder views, e.g. implementing partner, national partners as appropriate.
- The report clearly explains the extent to which the evidence supports the conclusions and judgments made.
- The evaluator makes their position clear. e.g. has the investment made adequate progress or not? Alternative points of view are considered appropriately.
- The conclusions and recommendations logically flow from the presentation of findings and any associated analyses.

The conclusions and recommendations logically flow from the presentation of findings and any associated analyses. It is possible to trace issues through the text from description, to analysis, to conclusion and recommendation. No recommendation appears at the end that is not supported by descriptive and analytical work in the text.

The “chain of evidence” is evident. This is where all questions in the methodology have data that has been collected, analysis conducted, findings presented, interpretation carried out and reported. If questions in the methodology have not been addressed then an explanation has been given.

Findings relevant to specific sub-groups (eg women, people with disability) are included.

The report makes it clear what issues are priority issues to consider. Minor issues are not set out mechanically against the terms of reference and given the same depth of treatment as more important issues.

The evaluator has made their position clear and the report presents their views unambiguously. For example, has the investment made adequate progress or not? Are the factors that have accounted for the limited achievements been unavoidable or are they due to poor management?

Alternative views are presented, especially for important, controversial or disappointing findings. They are not immediately dismissed, but are seriously considered. Key stakeholder views such as those of the implementation team must be given sufficient attention, and balanced by national partners, DFAT or other important stakeholder views.

Evaluator opinions that are based on limited evidence are made transparent and proposed as suggestive only.

Source: M&E Standards 6.6, 6.8, 6.9, 6.15, 6.16

8) Recommendations

Conclusions, recommendations and lessons are clear, relevant, targeted and actionable so that the evaluation can be used to achieve its intended learning and accountability objectives.

Any significant resource implications are estimated.

Major criteria:

- Recommendations are linked to significant findings, including lessons learned, emerging changes, opportunities or risks.
- Recommendations are clear, specific, relevant, targeted and actionable.
- Recommendations are realistic, i.e. likely to be effective to rectify a situation, or to achieve an expected outcome.

Findings and recommendations are feasible and, in the most part, are acceptable to relevant stakeholders. Recommendations are likely to be effective to rectify a situation, or to achieve an expected outcome.

Individuals have been allocated responsibility for responding to recommendations. Where appropriate, job titles, rather than organisations, have been allocated responsibility for actions against all recommendations. If it is not appropriate or possible to identify the individual, then the relevant work group is identified.

If recommendations imply human, financial or material costs, these are estimated.

Where there are important lessons to be learned, the report provides sufficient information to inform the reader about the circumstances under which these lessons can be transferred. This could be at the sector level, the country program level, for the Department as a whole, or for the development sector more broadly.

Source: M&E Standards 6.17, 6.18. 6.19, 6.20

9) Executive summary

The executive summary is standalone and provides all the necessary information to enable primary users to make good quality decisions

The executive summary provides all the necessary information to enable *primary stakeholders,* especially senior management, to make good quality decisions without reading the entire document.

It is not a simple cut and paste of the main body of the report.

It summarises the key findings, provides sufficient analyses and arguments, and presents final conclusions and recommendations.

Important information about gender equality and social inclusion are included to allow the reader to appreciate important achievements and challenges.

Resource implications of recommendations are summarised.

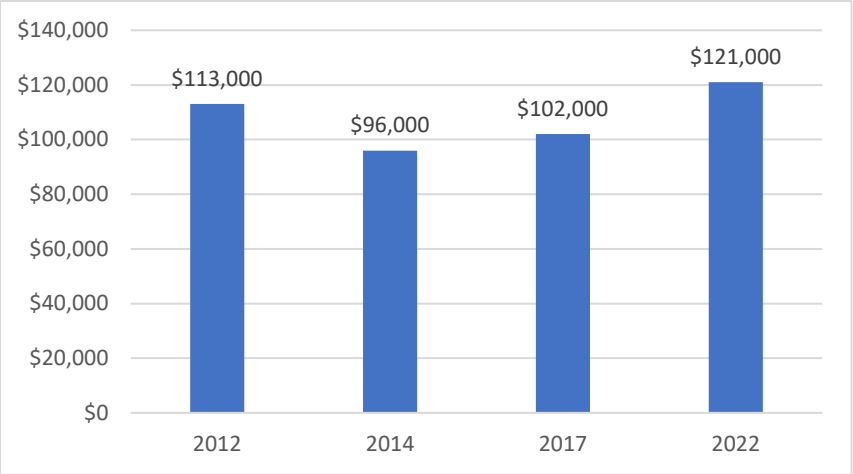
The length of the executive summary is proportionate to the length of the report (e.g. two to three pages for short uncomplicated reports, and up to five or six pages for more lengthy reports with complex issues).

Source: M&E Standard 6.4

ANNEX 4: AVERAGE EVALUATION COST OVER TIME

The average cost per evaluation has increased slightly over time. The average cost in 2022 was $121,164. This is an approximate cost for evaluation contracts and does not include the cost of DFAT staff time. Figure 7 compares this with the average costs from prior reviews (adjusted for inflation) and shows a slight increase over time. The median cost of evaluations has remained relatively modest at well below one percent of the cost of the investments in both 2017 and 2022.

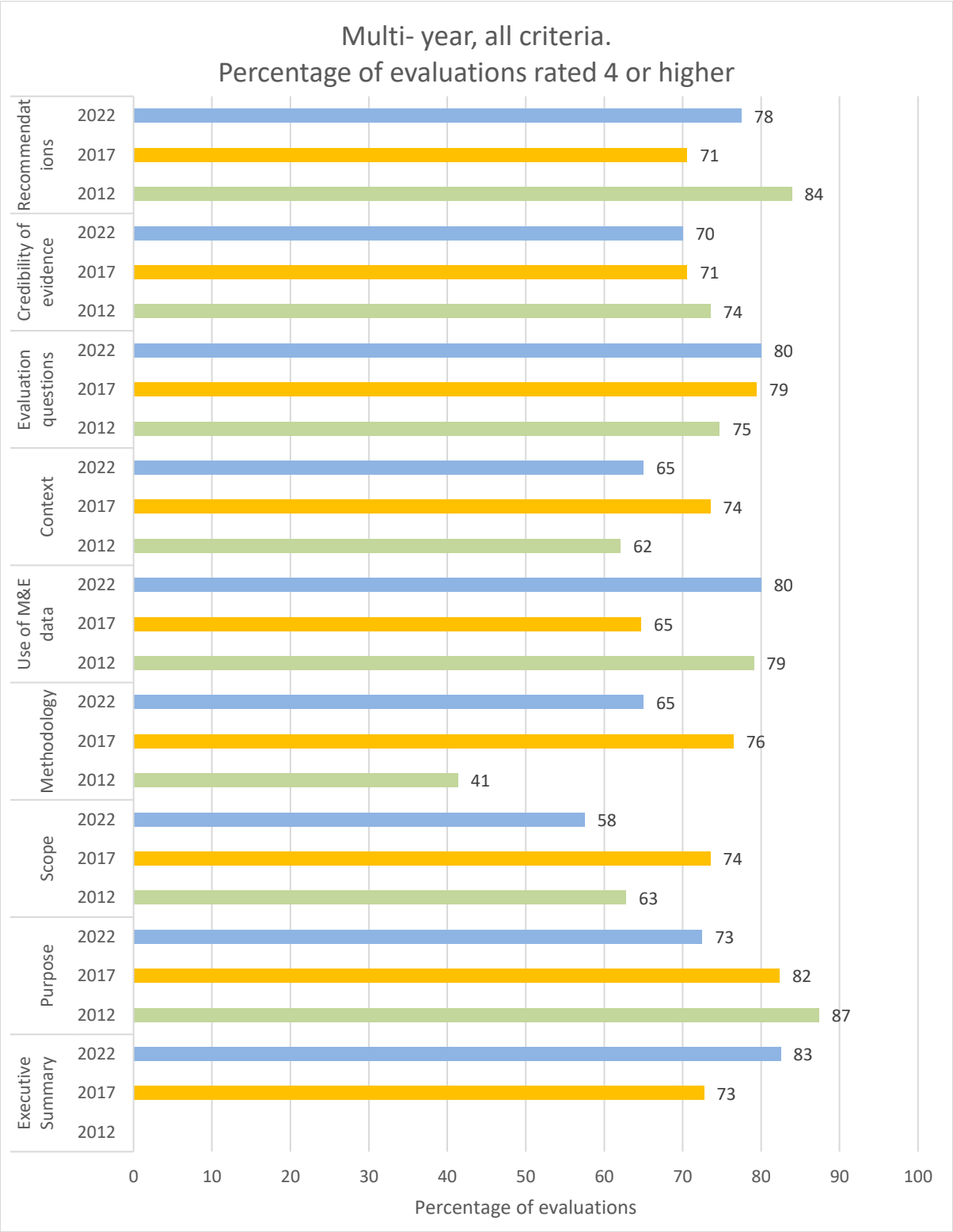
Figure 7: Average evaluation cost over time (adjusted for inflation and rounded)



The average number of contracted consultant days per evaluation was estimated at 79 per evaluation in 2022, an increase from 69 in 2017 and 72 in 2014.[19] In line with the costs, there has been a slight increase in the average number of consultant days.

[19] In many cases in all of these years, consultant contracting information had to be inferred from the available information as it was not directly accessible.

ANNEX 5: PERCENTAGE OF EVALUATIONS RATED ADEQUATE OR ABOVE FOR ALL QUALITY CRITERIA OVER TIME



Notes: Data for the review of 2014 evaluations is not included because it was not representative.

ANNEX 6: GOOD PRACTICE EXAMPLES

Name of evaluation	Aspects of good practice	Assessor comments
PNG Transport Sector Support Program Phase 2	Evaluation report	The evaluation report is clear with a logical layout. There are eight appropriate evaluation questions which are clear and specific, three related to program design and five to program implementation. The report is structured around the evaluation questions, which makes it easy to navigate. The authors clearly answer each question, bolding the key points that directly relate to answering the question. There is a description of the adequacy of the program's M&E system and some evidence that the report draws on the M&E system, particularly in relation to the achievement of outcomes. The conclusion summarises the key findings and provide useful context for the recommendations which directly follow each question. The methodology is appropriate, and the review draws on a range of sources with a coherent analysis of the evidence. The executive summary is clear and concise, providing information needed for future decision-making.
Australia, New Zealand, International Finance Corporation: Papua New Guinea Partnership	Evaluation report, TOR	The scope of this mid-term review (MTR), as detailed in the terms of reference (TOR), was well suited to the evaluation time and resources. The purpose described in the TOR was clear - to assess the progress of the partnership between Australia, New Zealand the International Finance Corporation in PNG towards promoting private sector development (PSD). The partnership was set up to support PSD in agriculture, tourism, power, financial markets, digital technology, public-private partnerships, and gender. The MTR examined the success and progress to date of underlying projects (both active and closed) and an assessment of the ongoing relevance of active projects given the impact of COVID-19. The evaluation used mixed methods to evaluate the partnership, incorporating quantitative and qualitative measures. The methods were appropriate for the task and explained well in the TOR. The MTR report was structured in relation to the six key evaluation questions and related sub-questions and addresses all questions in a succinct but comprehensive way. It identifies key strategic issues such as the lack of donor coordination, lack of attention to gender, and the impact of COVID-19 on activity implementation. The report also provides relevant operational and technical detail in the body of the report and in accompanying annexes. The evaluative judgements of the consultant are clear and backed by evidence. The report is not too long (28 pages for the main report plus annexes) and it is well written.
Justice Services and	Evaluation report,	The TOR for this evaluation were good, including a list of evaluation questions which were amended after discussion with

Stability for Development Program (JSS4D)	evaluation plan, TOR	the evaluation team. The scope of the evaluation was suitable for the time and resources available (90 days in total for the 3-person team). The skills and experience of the team, which included a gender specialist, were appropriate for the five key evaluation questions. The roles and responsibilities of the team members were clearly set out in the evaluation plan which was detailed and comprehensive, and appended to the main report. The evaluation plan also noted the responsibilities of the Australian High Commission and the Managing Contractor in coordinating and supporting the evaluation. There was time allocated to reviewing documentation and its appraisal. The evaluation plan included a table showing how each evaluation question would be answered in an evidence matrix with assessment criteria as the basis for analysis. The methodology described in the evaluation plan was very high quality. It described analysis and triangulation of data across the evidence matrix from a variety of sources to make a judgement of the strength of evidence. The evaluation used purposive sampling of stakeholders for interview, noting the option to use snowball or referral sampling where needed. The evaluation report included a table of possible limitations, and it was also clear on the limitations of evidence from the program MERL system. The evaluation plan included a section on safety and ethical practice which included cultural issues, confidentiality of interviewees, and also noted procedures for the safekeeping of data. Overall the report was well structured around the evaluation questions, each of which was clearly answered. The presentation of evidence is credible and convincing. Findings clearly flowed from the evidence and narrative. Where there were limitations in the data in specific areas this was made clear in the text. Noting the lack of evidence from the MERL system in assessing achievements against EOPOs, the evaluation team compared their own assessments of progress using documentary and stakeholder evidence with the program's assessment of progress to validate the program's reporting and assessments. A table summarises conclusions on progress, backed up with more detailed examination of progress against End of Program Outcomes and Intermediate Outcomes, with supporting evidence in an annex. Promoting gender equality was the focus of one evaluation questions and there are clear findings about the program's performance in this area. Disability is also considered.
Australian Infrastructure Financing Facility for Pacific	Evaluation report	This evaluation was very high quality in nearly every respect. The purpose and scope were very clear. The methodology was particularly sound. Of particular value was the use of an Evaluation Framework, which presented the evidence required to answer each key evaluation question, the data sources, the data collection methods and analysis approach. Document review and interviews were the primary sources as well as landscape analysis and benchmarking exercises. It was useful to see a comparison to

		other financing initiatives. Documents and interviews were coded against the key indicators required to answer the evaluation questions and NViVO software used to support the analysis. This is a good practice example of being clear about data analysis approaches and it is easy to see how data was intended to be triangulated. It is noted that certain factors were excluded where there was insufficient evidence to inform robust analysis. There was third party verification of the report and validation of the findings. Limitations were clearly identified. Ethical considerations were also included, covering informed consent and privacy and confidentiality. Other core strengths of the evaluation included clearly answering all of the evaluation questions and the credibility of the analysis in presenting the evidence. The recommendation section was also very high quality. The recommendations were explicitly linked to significant findings and prioritised. The four highest priority recommendations were to achieve long term strategic objectives, and six additional recommendations were to improve effectiveness and efficiency.
Penyediaan Air Minum dan Sanitasi Berbasis Masyarakat (PAMSIMAS)	Evaluation report	The report clearly set out the evidence for its conclusions and there was a clear line of sight along the chain of evidence. The evidence presented is credible and convincing, with key findings clearly backed up by data and sources referenced. The report contained a good level of evidence and supporting data and this was backed up by more detailed data (for example on water quality at the various sites tested) in the appendices. This was an example of an evaluation that made extensive use of the program's monitoring data throughout the report. It used the program's MIS data together with the primary data that the evaluation team collected to draw conclusions. It was notable that the evaluation interrogated the program MIS data where that data was not consistent with the evaluation's data collection and made recommendations for a future MIS. Evidence for the evaluation was thus collected from a range of sources including program MIS data, key informant interviews, focus group discussions, a small household survey, water quality testing, infrastructure quality checks. Data was triangulated to draw evidence-based conclusions. Conclusions flowed logically from the evidence presented. The appendices contained more detailed data that backed up the report's findings. WASH is a sector that lends itself to lots of quantitative data, but the report made a good job of using it and reporting on a very large program.

ASEAN Australia Smart Cities Trust Fund (AASCTF)	Evaluation report, TOR	This evaluation was very clearly structured so it is very easy to find information in the findings. The evaluation questions were appropriate and well answered, with a balance between operational and strategic issues. The conclusions are easy to read and flow well from the findings. A strength of the report was the executive summary, which clearly summarised the findings, conclusions and recommendations, with no significant information gaps. At three pages long it is proportionate to the length of the whole 36-page report. The ToRs are provided in the annexes and are a good practice example. The background is a good summary of a complicated program implemented in multiple locations. The evaluation questions are prioritised, and the sub questions are clear and relevant. The section on the requirements for the aide memoire is clear.
Independent Evaluation of the Australian NGO Cooperation Program (ANCP)	Evaluation report	The evaluation considers the issue of the Australian NGO Cooperation Program modality from a range of perspectives. The evaluation report answers each of the 5 key evaluation questions and the report is structured around these questions. The report comprehensively addresses the range of issues associated with each evaluation question. The literature review and comparative analysis of other donor practice ensures the evaluation has a broad view. Deep analysis is helpfully structured according to the conceptual frameworks used for analysis. The report summarises the findings in a conclusion chapter and then provides recommendations. This structure gives a clear line of sight between the findings and recommendations. The report draws out different perspectives well. For example, Australian NGOs and DFAT noted how accreditation contributed to good management systems and to recognition of them as good development partners. On the other hand, local development partners noted how these systems can bring about a focus on compliance while being useful for bringing about transformative change in relation to GEDSI. The evaluation's findings are based on strong evidence. That is, the key findings are triangulated across the multiple reliable sources (including the 2015 ODE evaluation, international literature, project documents and stakeholder consultations). Where there is a lack of evidence, this is noted. There was a range of formats used to present data, particularly in the annexes.