REVIEW OF 2017 PROGRAM EVALUATIONS - ANNEXES

ANNEX A: EVALUATION PLAN

1. Introduction

The Office of Development Effectiveness (ODE) will conduct the Review of Program Evaluations completed in 2017 (the Review). This will examine the quality of independent program evaluations completed in 2017 and facilitate opportunities to learn from program evaluations. The Review follows on from the previous two Reviews of Program Evaluations, which examined program evaluations completed in 2012 and 2014.

2. Background

Independent evaluations are undertaken at two levels in the department:

- *Strategic evaluations* are produced by ODE. These are high-level evaluations of aid program policies, strategies and approaches to common development issues

- *Program evaluations* are managed by country and regional programs. These focus on evaluations of individual/clusters of aid investments or thematic evaluations at a program level.

ODE has completed two previous Reviews of Program Evaluations: (1) a review of 87 independent program operations completed in 2012 (all evaluations) and (2) a review of 35 independent evaluations completed in 2014 (purposive sample).

Since the second Review, DFAT has revised its evaluation policy. The new aid evaluation policy, introduced in late 2016, aims to ensure a "demand driven" and "fit for purpose" approach to evaluation by providing programs with the flexibility to determine the highest priority issues their evaluations should focus on. Requirements under the revised evaluation plan include: (1) programs are given a minimum number of evaluations which should be conducted each year with larger programs expected to undertake more evaluations; and (2) each year ODE compiles DFAT's Annual Evaluation Plan which will be reviewed and approved by the Secretary and shared with the Minister for Foreign Affairs.

The new evaluation policy has been implemented successfully to date. DFAT's first annual evaluation plan under the new aid evaluation policy was developed in 2017. At the end of 2017, 41 out of 43 evaluations on the revised evaluation plan were published with management responses. Prior to the new policy, only half of DFAT's aid evaluations had management responses and only a third were published (2016 ODE Review of Evaluations).

3. Rationale

The Review is proposed for the following reasons:

Firstly, the review provides an opportunity to assess the impact of the current Evaluation Policy on evaluation practice, quality and use. The introduction of the new evaluation policy in November 2016 improved the publication rate of evaluations and preparation of management responses significantly but we have no information about their quality. By comparing program evaluations conducted in 2017 under the new aid evaluation policy with those assessed by the previous Reviews, the Review provides the opportunity for insights into how the changes to evaluation policy have affected the quality and use of program evaluations.

Secondly, the review will give staff (senior managers, investment managers, performance and quality staff, and ODE) a stronger understanding of current program evaluation practices, and how further improvements can be made. This will include identifying the features of high quality program evaluations, for example evaluation resourcing, evaluation design, and organisational factors needed.

Thirdly, the review will ensure DFAT understands how credible program evaluations are as a source of performance information for the Australian aid program.

Finally, the review will provide an opportunity for ODE to identify and disseminate the key lessons from evaluations for the Department and the broader aid community.

4. Objectives

The Review has three objectives:

a. 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 and 2015.

b. To provide information to support good quality, independent evaluations across the department; and

c. To promote better use of evaluations across the department and the aid community by extracting, synthesising and disseminating valuable insights and lessons from evaluations to both DFAT and the broader aid community.

5. Scope

The Review will examine all the 37 independent program evaluations completed in 2017. This includes evaluations commissioned by DFAT as well as joint or partner-led evaluations. The 2017 evaluations will give the most up to date data on current evaluation practice in the department.

Reviewing all 37 evaluations will allow us to draw firm conclusions that apply to all program evaluations completed in 2017. A full sample will also provide insight on whether the new evaluation policy has resulted in evaluations which provide adequate coverage of the aid program (in terms of sector, geographic focus, and funding).

6. Audience

There are two primary audiences for Phase 1 of this Review (quality review of evaluations): staff from ODE and performance and quality (P&Q) staff from DFAT's program areas. The most relevant evaluation questions for both of these groups, and how the evaluation findings can be used by them, are outlined in Annex A.

The key secondary audiences for Phase 1 of the Review are the DFAT Executive and senior managers from the Contracting and Aid Management Division (ACD). The findings from the Review can be used by these groups to inform decisions on the department's investment quality reporting system, investment designs and implementation, and to inform DFAT M&E training.

The main audiences for Phase 2 (synthesis of learning) of the Review include senior DFAT officers at Post and in Canberra and the broader aid community.

All DFAT staff involved in commissioning and managing evaluations will also have an interest in the Review's findings in assisting them to commission and manage higher quality evaluations and in the opportunities for learning which are provided by the Review which may inform program design, implementation and management.

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

› The Review report will be published on the ODE website

› A four-page summary will be printed and distributed at appropriate forums (see below)

› ODE staff and consultant will present the Review process and findings at appropriate DFAT and other forums as opportunities arise. e.g. ODE-ANU Evaluation Forum, Australasian Evaluation Society's annual conference, Annual Australasian Aid Conference

7. Approach

The Review will be conducted in two phases:

a) **Phase 1 (April – September 2018)** – Quality review of program evaluations conducted in 2017. All the 37 independent program evaluations completed in 2017 will be quality reviewed.

b) **Phase 2 (July – November 2018)** – Synthesis and dissemination of lessons from 2017 program evaluations. The synthesis of learning will focus on two current areas of interest to the aid program: policy influencing and aid capability, as well as any other emerging themes identified in the 2017 program evaluations.

The review of the quality of 2017 evaluations (Phase 1) will be conducted by a team of six, including five ODE staff and a consultant, who will assist with the synthesis of learning component of the Review. The consultant will be a member of the evaluation team and will also assess the quality of six program evaluations.

Conducting the evaluation largely in-house will promote retention of learnings from the evaluation by the department and build internal knowledge, analytical and evaluation skills. It will also contribute to an efficient process and useful findings which are appropriately targeted.

The Review of 2014 evaluations, which was comparable in scale (35 evaluations reviewed compared to 37 evaluations proposed in 2017), was successfully conducted internally with a team of six ODE staff using a reasonable investment of staff time. ODE currently has the capacity to undertake the Review in-house following the recent recruitment of five permanent staff into vacant positions. Drawing on records of ODE staff time taken for the 2015 Review it is estimated that the review will take approximately 6 months to complete.

8. Evaluation Questions

The following evaluation questions define the scope of the evaluation and will guide the data collection and analysis undertaken by the evaluation team.

Priority questions

1. What are the characteristics and quality of program evaluations? How have these changed over time?

2. What factors contribute to the quality of program evaluations?

3. To what degree do program evaluations a credible source of evidence for the effectiveness of the Australian aid program?

4. What can be learned from the evaluations, particularly in the areas of policy influence, aid capability and gender equality about how context [1]affects outcomes and implications for DFAT?

---

[1] Context refers to aspects of the particular situation before the intervention. It can include any characteristic such as culture, norms, socio-economic factors of individuals, institutions and the broader environment.

Other key questions

5. How are the findings from program evaluations used in the Department?

6. Based on the findings of this review, what are the implications for the department's evaluation policy?

7. Which evaluations can be nominated for a Secretary's award for evaluation excellence?

9. Methods

Attachment 1 sets out the data sources, collection methods and tools which will be used to gather relevant evidence against each of the evaluation questions.

Assessing evaluation quality and use

Phase 1 will largely focus on assessing the quality of program evaluations completed in 2017. The method for assessing evaluation quality is a desk review of all program evaluations published in 2017. The quality of each program evaluation will be assessed by an adaption of the same assessment pro-forma used in previous Reviews, which draws on nine criteria based on DFAT's Monitoring and Evaluation Standards. The nine quality criteria include: 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. The assessment results will be recorded in a central database and each criterion will be analysed to identify specific areas where evaluation quality is high or low.

A similar approach for assessing evaluation quality was successfully used in the previous Reviews. Using this same method will ensure that evaluation quality and practice can be compared to the 2012 and 2015 Reviews to measure any changes over time and under different evaluation policies. The assessment pro-forma will be reviewed by the evaluation team prior to the evaluation and adapted if necessary to meet the specific requirements of the Review. See Appendix 2 for a copy of the assessment pro-forma.

A measure for overall evaluation quality will be established. The criterion "credibility of evidence and analysis" was used as a proxy for overall evaluation quality in the 2012 and 2015 reviews as the criterion was most strongly associated with other quality criteria in the reviews. For this Review, correlation analysis of the nine pro-forma criteria will be undertaken to establish whether "credibility of evidence and analysis is still the best predictor of evaluation quality. If so, this criterion will continue to be used as a proxy for evaluation quality.

Factors contributing to quality

As in previous reviews, factors contributing to the quality of evaluations will be examined and analysed to identify which factors make the strongest contribution to evaluation quality. Factors to be considered will include: evaluation duration, purpose, team size, team composition, quality of M&E systems and number of evaluation questions. Correlation analysis will be conducted to examine the relationship between the identified factors and evaluation quality. Given the small number of evaluations that will be examined, the Review will need to treat the findings cautiously when making inferences. However, the results of the analysis, combined with the findings of other meta-evaluations, will assist DFAT to identify a small number of critical areas that can contribute to high quality evaluations.

Note that a number of other factors that contribute to evaluation quality have been identified through other meta-evaluations. These include the quality of the evaluation design; the presence of a strong relationship between the commissioning agency and the evaluation team; and the evaluation capacity of staff in the commissioning agency. However, the relationship between such factors and evaluation quality will not be examined in this Review. This is because significant resources would be needed to collect and analyse the qualitative data required to understand these factors, which are beyond the capacity of the Review team. It is

possible that other factors that contribute to evaluation quality will emerge during the Review. If team members identify such factors, they will be discussed among the team to determine whether they should be further examined.

Investment and evaluation characteristics, quality and factors contributing evaluations will be compared to the results of the previous Review through a series of tables and figures. Analysis will focus on areas where the characteristics or quality have changed substantially since the previous Review.

The degree to which program evaluations provide a credible source of evidence for the effectiveness of the Australian aid program will be addressed by examining the assessments against pro-forma criterion 8, 'credibility of evidence and analysis'. Note that, given conclusions will be drawn from this single source of evidence, their strength will be modest.

Use of program evaluations

The use of program evaluations will be assessed by looking at management responses for evaluation reports to determine the number and percentage of evaluation recommendations which have been accepted, partially accepted, or not accepted.

Evaluation use will also be assessed by distributing a small survey to relevant staff in Canberra and Post to identify how evaluation recommendations are being used to influence policy and program development and constraints to implementing evaluation recommendations.

Implications for DFAT's evaluation policies and practices

To ensure the Review provides information to support good quality independent evaluations across the department, the findings from across the evaluation questions will be brought together to identify the implications for DFAT's evaluation policies and practices. Analysis will focus on whether DFAT's current evaluation policy is encouraging intended practices in the department, potential areas where DFAT's evaluation policy could be adjusted, and lessons that could be adopted from other donors.

Opportunities for Learning

Data identification and extraction

The consultant will have main responsibility for the data extraction and synthesis of learning from the program evaluations.

The synthesis of learning on policy influencing and aid capability will be conducted by testing a new approach to synthesis that examines the ways in which different contexts affect the outcomes of activities by identifying the mechanisms that are interacting with the different contexts to produce the outcomes. This approach – a realist approach [2]- has a growing body of interest globally among evaluation practitioners, but relatively few examples in practice.

The consultant has drawn on relevant literature on realist synthesis to develop a methodology for the extraction of data and the synthesis. The consultant will brief the ODE team on realist synthesis and how to extract data from the evaluation reports.

---

[2] "Realist synthesis is underpinned by a 'generative' understanding of causation, which holds that, to infer a causal outcome/relationship between an intervention (e.g., a training program) and an outcome (O) of interest (e.g., employment), one needs to understand the underlying mechanism (M) that connect them and the context (C) in which the relationship occurs (e.g., the characteristic of both the subjects and the program environment - culture, norms, socio economic factors etc). The interest of this approach is not simply which interventions work, but which mechanisms work in which context. Rather than identifying replications of the same intervention, the researcher should adopt an investigative stance and identify different contexts in which the same underlying mechanism is operating. Realist synthesis is concerned with hypothesizing, testing and refining such context-mechanism-outcome (CMO) configurations. Based on the premise that program work in limited circumstances, the discovery of these conditions become the main task of realist synthesis. The overall intention is to first create an abstract model (based on the CMO configurations) of how and why program work and then test this empirically against the research evidence. Thus, the unit of analysis is the program mechanism, and this mechanism becomes the basis of the search. This means that a realist synthesis is not defined by topical boundaries and can range across a wide area, but its search aims to identify different situations in which the same program mechanism has been attempted." (Gough et al, p.54)

Phase 1 will include extraction of data from the program evaluations for the synthesis in Phase 2. For each evaluation report and any associated documents, the Review Team will be looking for two things:

(1) *Explanatory text:* that is an explanation of how particular intended outcomes came about or did not come about as a result of the intervention and context. The Review Team will select the section of explanatory text and insert a comment that includes the relevant codes as specified the pro-forma and guidance notes:

(2) *Policy influence:* if policy influence was involved in the initiative(s) being evaluated then the Review Team will identify the approach or approaches being used (Advising, Advocacy, Lobbying. Activism, Other approach) and also identify if Gaming seemed to be involved (this is where the country 'pretends' to have their policy influenced to please the donor) and specify these in the pro-forma. Text does not need to be selected for this, just a yes in the pro-forma.

The assessment pro-forma used to conduct the quality review of evaluations will also be the tool for collating this information. The 'synthesis' section in the assessment pro-forma includes a space to identify which evaluation reports contain which of the coded information, and which approach to policy influence, if applicable.

The coded documents will be shared with the consultant on a progressive basis during Phase 1.

The consultant will provide ongoing technical support to the ODE team during data extraction from the evaluation reports in Phase 1.

Ongoing modification, interviews (if necessary) and synthesis

The data extraction process will be done progressively, with coded material periodically copied into a consolidated table by the consultant and reviewed and analysed by the consultant. Any variations needed to the method will be discussed and modifications made, if necessary, during the regular moderation meetings.

It is anticipated that clear detail on explanations may not be well documented. Specific areas where more information is needed will be identified. Possible sources of the information will be identified (DFAT project managers or evaluators) and, if agreed by ODE, interviews undertaken by the consultant. Up to 20 interviews may be undertaken, most likely by telephone. Transcripts of interviews will be coded in the same way as described above.

The consultant is responsible for the synthesis of the extracted data. The synthesis will be guided by the evaluation questions. It will commence in the early stages of data extraction and iteratively return to the literature, analyse evidence from the evaluations and inform the direction for the ongoing data extraction and possible interviews.

The consultant will discuss the findings with ODE and a decision will be taken on the type and number of 3 to 5 short learning products for dissemination in DFAT and the broader aid community. These will be drafted, reviewed by ODE and finalised by the consultant.

Other learning opportunities

We will also look for opportunities to disseminate evaluations that are high quality and highly relevant. For example, if the Review identifies evaluations which meet a pre-determined quality threshold and which are relevant to a strategic evaluation, these could be used to inform that strategic evaluation. Other high quality evaluations can also be brought to the attention of areas of the department that may have an interest in the findings (for example, governance, health, environment, fragile states, etc). These areas of the department can decide on the best use for the program evaluations; for example, they could complete their own synthesis or use them to inform their ongoing work in policy development, program design and research. ODE would not necessarily have ongoing involvement in this work.

10. Resources

As noted previously, the Review will be conducted by a team of six, including five ODE staff with the assistance of a consultant for the extraction and synthesis of learning. Table 2 outlines the team and roles.

All team members will be asked to record the approximate time they spend working on the Review. This will allow ODE to better understand the staff time and resources required to implement the Review.

A budget of approximately $40,000 is required for the Review. This will include funds to contract the consultant to undertake the extraction and synthesis of learning (up to $37,500) and for editing and printing services (up to $2,500).

Table 1: Review roles and responsibilities

Review position	ODE position	Key roles in Review
Team Leader	Director, Program Evaluations Section	Accountable for the Review. Tasks include: › Overseeing Review process › Ensuring consistency of assessments across the team › Assessing the quality of program evaluations
Review Manager	Assistant Director, Program Evaluations Section	Coordinate day to day Review tasks, including: › Creating data recording and management systems › Assessing the quality of program evaluations › Coordinating team members' input › Conducting data analysis › Coordinating and drafting key documents with input from other review team members
Review Assistant	Policy Officer, Program Evaluation Section	Responsible for data management, including: › Collecting program evaluation documents › Recording basic characteristics of evaluation › Data entry › Research tasks as needed
Other team members		› Assessing the quality of program evaluations › Contributing to data analysis and drafting key documents › Assistant Director OES will manage data collection and analysis for evaluation question 5
Consultant		› Design methodology for the synthesis of learning › Brief and assist team with data identification and extraction › Assess the quality of program evaluations › Contribute to data analysis and drafting key documents › Undertake synthesis

11. Limitations

The terms of reference identified that a limitation is that ODE will be assessing the evaluation policy, guidance and support which it oversees and provides. This limitation will be addressed as follows:

› The Independent Evaluation Committee (IEC) will oversee the Review process and findings. This will help ensure any self-assessment conducted by ODE is defensible, and

› The limitation will be clearly acknowledged in the evaluation report to ensure readers take it into account.

Another limitation identified in the terms of reference is that ODE will likely be required to address the Review's findings and complete a management response. It will be difficult for ODE to draft report recommendations that ODE will then have to respond to. Because of this the Review report will not include recommendations; rather, it will outline findings and then note the steps ODE will take to address these findings.

A structured search would aim to identify all evaluations of direct relevance to evaluation question 4. However, the set of evaluations is pre-selected here is limited to 37. This means that the synthesis will not be comprehensive, rather it will aim to find new insights for DFAT and the broader aid community from the available data but this will always be strengthened through further work. The consultant will mitigate as far as possible within the constraints by further reference to the literature.

Consistency of assessments across the team

The quality of program evaluation reports will be assessed by up to five team members and the consultant using the pro-forma 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. This will be achieved as follows:

a) A Review handbook will be created which provides a short description of each criterion, including what 'adequate' quality for each criterion looks like.

b) Prior to starting assessment of program evaluations, the Review team will meet and discuss the assessment pro-forma and handbook. This will provide a common understanding of the criteria and the ratings to be used.

c) As the first step in the assessment process, all team members will assess the same program evaluation using the pro-forma. The Review team will then meet and discuss the assessments (moderation session). This will further build a common understanding of how to assess the evaluations and identify any criteria where further clarification or changes to the pro-forma are required.

d) If the team leader believes further work is required to ensure team members provide relatively consistent assessments, Step 2 will be repeated with a second program evaluation.

12. Ethical conduct

Consistent with ODE policy, the Review will be guided by the Australasian Evaluation Society (AES) Guidelines for the Ethical Conduct of Evaluations. The consultant is a member of the AES. The consultant is also bound by the ethics approval for their research at ANU (protocol 2018/052). This requires the consultant to obtain informed consent from interviewees if interviews proceed.

Issues around ODE's assessment of its own policies, guidance and evaluations are discussed under 'Limitations' above.

13. Schedule

The Review tasks to be undertaken, persons responsible, the approximate time required for each task, and approximate dates for task completion are in Table 3. *Replace Table below with revised version I sent out to evaluation team*

Table 2: Key evaluation activities and schedule

Task	Person(s) responsible	Working days required	Dates for completion
Evaluation planning			
Develop evaluation plan.	Review Manager	10 days	Final draft by end of April
Collect evaluation documents (including reports, ToRs, evaluation plans and management responses).	Review Assistant	37 evaluations x 45 minutes each Total: 3.5 days	Completed
Set up evaluation systems (eg data management systems).	Review Manager and Assistant	4 days	4 May
Review inception meeting	Review Manager (all team members to attend)	1.5 hour per team member (9 hours total)	7 May
Data collection			
Assess the quality and credibility of each operational evaluation (including ToRs and evaluation plans, where available) using the Review pro-forma. This process will include: • Identifying 4-6 good practice examples • Identifying, coding and recording of potential learning information • Extracting learning information	Review team – identifying good practice examples, identifying, coding and recording learning information The consultant will have primary responsibility for extracting relevant learning information	Quality assessment: 37 evaluations x 7 hours each Total: 37 days	14 May – 13 June
Moderation of evaluation to strengthen assessment consistency.	Review team	2 evaluations (14 hours per team member) + 2 moderation meetings of 1 hour each Total per team member: 2.1 days (13 days for a team of 6)	First moderation meeting 14 May Second moderation meeting 28 May
Data on evaluation characteristics and quality to be recorded in central location.	Review assistant	36 evaluations x 30 minutes each Total: 2.5 days	14 May to 13 June

Create and distribute a survey to gather data on the department's use of evaluation recommendations to influence policy and programming	Review Manager	1 day	14 June
Data analysis			
Quantitative analysis to identify average characteristics of operational evaluations (for example, average cost; average team size etc)	Review Manager, with assistance from team members	4.5 days	15 June – 29 June
Quantitative analysis to identify the relationships between such characteristics and evaluation quality.	Review manager, with assistance from team members	3.5 days	15 June – 29 June
Analysis of results from the survey on evaluation use	Assistant Director OES	2 days	15 June – 29 June
Synthesis of extracted data	Consultant	5 days	15 June – 29 June
Team workshop to discuss analysis and agree on key findings of quality review	All team members	1/2 day per team member (3 days for team of 6)	4 July
Report writing and review			
Draft report on quality review	Review Manager to coordinate team members	13 days	5 July – 3 August
Review processes, including IEC review and report revisions.	Review Manager	20 days	6 August – 14 September
Draft, revise and finalise 3 – 5 learning products, including IEC review	Consultant	20 days	August – November (consultant will be away for some of this time)
Total days required			

14. Outputs

Phase 1 outputs will include:

- An evaluation plan outlining the detailed methods to be used for the evaluation.

- An evaluation report outlining the key findings of the quality review the program evaluations completed in 2017, including recommendations for further synthesis/learning products on policy influence, aid capability and other themes that can be developed from key lessons extracted from the 2017 evaluations

- A collection of 4-6 good practice evaluation products, including evaluation reports, ToRs, evaluation plans and management responses. This will include identification of exemplar evaluations that can be nominated for the Secretary's Award.

- Lists of relevant program evaluations for thematic areas of the department.

- Detailed records of data collected and analysis undertaken. This will be retained for ODE records, with relevant parts included in the evaluation report.

- Report of a reflection exercise – what worked, what didn't work, lessons learned – from the Review process, which can be used as a resource for planning and implementing the next RPE.

The outputs in Phase 2 will be informed by the findings from Phase 1 but the synthesis could include 3 - 5 synthesis briefs /key learnings about policy influence, aid capabilities and any other possible emerging themes.

ATTACHMENT 1: Summary of evaluation questions and methods

Evaluation question	Primary intended users	Data collection methods	Data analysis methods
1. What are the characteristics and quality of program evaluations? How have these changed since 2012 and 2014?	ODE: to better understand current practices in the department and to provide relevant advice to the Executive	Basic characteristics of 2017 evaluations (and the investments they relate to) collected from evaluation database, AidWorks and Review pro-forma 2017 evaluations rated against quality criteria in Review pro-forma Evaluation characteristics and quality summarised from the 2012 and 2014 Review of Program Evaluations	Descriptive statistics to be derived, such as number, average and range for evaluation and investment characteristics Analyse each pro-forma criteria to establish areas where evaluation quality is high and low. Establish a measure to assess overall evaluation quality. Comparative tables/ figures of evaluation characteristics and quality, comparing with previous two Reviews of Program Evaluations Analysis of comparatives tables to assess changes to evaluation practices.
2. What factors contribute to the quality of program evaluations?	ODE, ACD and P&Q staff: to provide advicetraining to program staff on factors to focus on when planning evaluations	Data on such factors and evaluation quality collected under Q1 above. Data on quality of investment M&E systems to be collected from Aid Quality Reports.	Correlation analysis to examine relationship between evaluation quality and possible factors contributing to evaluation quality
3. To what degree do program evaluations provide a credible source of evidence for the effectiveness of the Australian aid program?	ODE staff: to support ODE assessments of the Investment Quality System and the Performance of Australian Aid report	Data on evaluation quality collected under Q1 above.	Analysis of assessments against pro-forma criterion 8, 'credibility of evidence and analysis'.
4. How are the findings from program evaluations used in the department?	ODE: to better understand current practices in the department and to provide relevant advice to the Executive	Management responses assessed to examine which recommendations accepted Short survey to relevant DFAT staff in Canberra and at Post to identify: (1) how evaluation recommendations are being used to influence policy and program development; and (2) constraints to implementing evaluation recommendations.	Examine percentage of recommendations which were fully, partially or not accepted Analysis of how evaluations are being used to influence policy and program development. Analysis of factors that are hindering the uptake of evaluation recommendations.
5. Which evaluations can be nominated for a Secretary's award for evaluation excellence?	DFAT staff: to highlight and incentivise best practice in evaluation.	The top five ranked evaluations against quality criteria in the assessment pro-forma will be nominated for the Secretary's award. Evaluations must also have a score of at least 5 for each of the nine quality criteria in the assessment pro-	Nominated evaluations will be assessed against an additional set of criteria. This could include some of the following: - evidence that the evaluation focussed on the highest priority issue for the country program/thematic area

		forma to be eligible for nomination for the Award.	- the evaluation examined an innovative investment; - evidence that the evaluation was used to inform future programming; - valuable lessons addressing priority issues for the aid program were able to be extracted from the evaluation; - one or more of the evaluation's other documentation (e.g. TORs; evaluation plan; management response) was identified as a good practice example; - use of innovate evaluation methods; - evidence of achievement of higher order outcomes or evaluation led to a significant change in the investment or department. ODE may require to follow up program areas to gather further information in order to fully assess nominated evaluations against these additional criteria (e.g. evaluation informed future programming).
6. Based on the findings of this review, what are the implications for the department's evaluation policy?	ODE: to better understand the impact of the current Aid Evaluation Plan (2016) on evaluation practice and use and identify potential areas where the policy could be adjusted.	Data and analysis from Q1, Q2, Q3 and Q4 will be collated to identify the implications for DFAT's evaluations policies and practices.	Analysis will focus on whether DFAT's current evaluation policy is encouraging intended practices in the department and potential areas where DFAT's evaluation policy could be adjusted.
7. What can be learnt from the evaluations, particularly in the areas of policy influence, aid capability and gender equality about how context affects outcomes and the implications for DFAT?	DFAT senior and program staff at Post and in Canberra to maximise the use of program evaluation findings to inform aid programming.	The Review Team will identify and code sections of explanatory text in the evaluation reports that relate to different outcomes and to aspects of aid capability. They will also identify which evaluations relate to specific approaches to policy influence, if applicable	Analysis will focus on identifying the same mechanism operating in different situations and how the context and intervention affect the outcomes achieved

ANNEX B: PRO-FORMA FOR ASSESSING THE QUALITY OF PROGRAM EVALUATIONSCOVER SHEET

The cover sheet is divided into two columns:

- In the first column, titled 'Numerical or Yes/No data', only enter numbers single letter answers (Y, N etc). This is so data from this column can be easily sorted
- In the second column, additional notes and information can be provided if needed.

Investment value	Insert value of investment in the first column if it is stated in the report
Cluster evaluation?	In the first column, record: » 'Y' if it is a cluster evaluation (ie it evaluates more than one DFAT investment) » 'N' if it is not a cluster evaluation (ie it only evaluates one DFAT investment)
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 second column outline the partner(s) involved in the evaluation.
Evaluation cost (if available)	In the first column, record the evaluation cost (eg $105,000) if it is available in any of the evaluation documents. Note we expect many evaluations will not include this information.
Number of evaluation team members	In the first column, record the total number of evaluation team members as outlined in evaluation documents.

	This should include people with substantive roles in data collection and analysis and/or report writing. It should **not** include observers who participate in field visits. If different evaluation documents (ToRs, evaluation plan and evaluation report) give different information on team size, use the information from the most recent evaluation document.
Evaluation team leader skills: specialist evaluation skills?	In the first column, record: » 'Y' if the team leader is a specialist evaluator » 'N' if the team leader is not a specialist evaluator. If you record 'N', in the second column outline the main skills of the team leader.
Evaluation team skills: main skills of evaluation team members	Do not record anything in the first column (it has been greyed out). In the second column, record the main skills of team members, as outlined in the evaluation plan and/or report. These could include evaluation skills, technical/sector skills and country/regional knowledge. If the composition and skills of the evaluation team are not clear, record "unclear" in the second column.
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 second 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 second column.
Evaluation duration - fieldwork days	In the first column, record the total number of fieldwork days for all team members. For example, if there are 2 team members and they are in the field together for 14 days, record 28 as the total number of fieldwork days. Fieldwork days include all working days outside Australia, including travel days. They do <u>not</u> include data collection in Australia. In the second column, provide the breakdown of fieldwork days per team member. For the example above, record: » Team leader: 14 days » Team member: 14 days If fieldwork days are not available or are unclear, record 'not available' in the second column.
Evaluation duration - total person-days	In the first column, record the total number of work days for all team members. For example, if the team leader works on the evaluation for 40 days and a team member works for 35 days, record 75 as the total number of person days. Focus on the number of days team members worked on the evaluation, rather than the total date range. For example, if an evaluator spent three days drafting a report over a 10 day period, record three days.

	In the second column, provide the breakdown of total days per team member. For the example above, record: » Team leader: 40 days » Team member: 35 days If the evaluation duration is not available or is unclear, record 'not available' in the second column.
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.
Performance criteria assessed?	In the first column, record: » 'Y' if at least one of the DFAT performance criteria (eg relevance, effectiveness, efficiency etc) is assessed » 'N' if no DFAT performance criteria are assessed If you record 'Y', in the second column record which criteria are assessed and page references.
Numerical ratings for performance criteria provided?	In the first column, record: » 'NA' if you answered 'N' to the question above » 'Y' if at least one of the DFAT criteria receives a numerical rating from 1-6 » 'N' if the DFAT criteria do not receive a numerical rating If you record 'Y', in the second column outline which criteria received a numerical rating and provide page references.
Number of recommendations in the evaluation report	In the first column, record the number of recommendations in the evaluation report
Management response located?	In the first column, record: » 'Y' if the management response is available » 'N' if the management response is not available
If no management response, are management recommendations followed up through other means (e.g. working groups)?	In the first column, record: » NA if you answered 'Y' to the question about » 'Y' if recommendations will be followed up through means other than a management response » 'N' if it is not apparent how recommendations will be followed up If you 'Y', in the second column outline how recommendations will be followed up, for example through working groups, working planning processes etc.
Management response: number of recommendations accepted?	In the first column, record the number of recommendations that have been accepted in the management response.
Management response: number of	In the first column, record the number of recommendations that have been partially accepted in the management response.

recommendations partially accepted?	In the second column list the key reasons why recommendations are only partially accepted.
Management response: number of recommendations not accepted?	In the first column, record the number of recommendations that have been accepted in the management response. In the second column list the key reasons why recommendations are not accepted.

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

The detailed descriptions of the 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 C: PROGRAM EVALUATIONS REVIEWED

Investment number	Evaluation title	Country	Primary sector
INJ577	Vietnam: Integrated Coastal Management Program	Vietnam	Agriculture, Fisheries and Water
INK781	Cambodia Agricultural Value Chain Phase 1 - undertaken by ODE	Cambodia	Agriculture, Fisheries and Water
INL412	Greater Mekong Water Resources Program	Multi-country	Agriculture, Fisheries and Water
INK103	Sri Lanka: Community Forestry Program	Sri Lanka	Agriculture, Fisheries and Water
INK130	Fiji Community Development program	Fiji	Community Development
INL451	Bougainville Youth Initiative	PNG	Community Engagement
INK784	Timor-Leste: National Program for Village Development Support Program	Timor Leste	Community Engagement
INL343	Philippines: Australia's Support for Peace in Mindanao	Philippines	Community Engagement
INK227	Australia Afghanistan Community Resilience Scheme	Afghanistan	Community Engagement
INI898	Pacific Financial Inclusion Program	Multi-country	Economic Development
INK403	Cleared Ground De-Mining Project in Palau	Palau	Economic Development
INJ498	Australia Indonesia Partnership Rural Economic Development Program	Indonesia	Economic Development
INL277	Mekong Business Initiative	Multi-country	Economic Development
INL833	Pakistan Trade and Investment Policy Program	Pakistan	Economic Development
INK496	Pacific Women Shaping Pacific Development	Multi-country	Gender Equality
INI402	Pacific Women Shaping Pacific Development: activities in PNG	PNG	Gender Equality

INL397	Fiji: Pacific Women's Country Plan Review	Fiji	Gender Equality
INL236	Pakistan Ending Violence Against Women Program	Pakistan	Gender Equality
INL307	Independent Evaluation of the Pacific Leadership Program Phase 3	Multi-country	Governance
INK455	Vanuatu: Governance for Growth	Vanuatu	Governance
INJ858	PNG Health and HIV Multilateral Partnership	PNG	Health and Education
INI456	Samoa Inclusive Education Demonstration Program	Samoa	Health and Education
INK372	Vanuatu: Strengthening Early Childhood Care and Education	Vanuatu	Health and Education
INJ648	Australia's Education Partnership with Indonesia	Indonesia	Health and Education
INK910	Australian support for improved nutrition in Timor-Leste	Timor Leste	Health and Education
68903	Strengthening Pre-service teacher training in Myanmar	Myanmar	Health and Education
INL754	Australia's response to Nepal earthquakes	Nepal	Humanitarian
INM002	Tropical Cyclone Winston Education Response	Fiji	Humanitarian
INL555	Australia's humanitarian assistance to Myanmar	Myanmar	Humanitarian
INL847	Australia response to El Nino in PNG	PNG	Humanitarian
INL435	Cambodia: 3i – Investing in Infrastructure	Cambodia	Infrastructure/ Transport
INK771	Vanuatu: Roads for Development	Vanuatu	Infrastructure/Transport
ING406	Eastern Indonesia National Road Improvement Project - undertaken by ODE	Indonesia	Infrastructure/Transport
INH582	Indonesia Infrastructure Initiative Phase 2	Indonesia	Infrastructure/Transport