Australian Government

Department of Foreign Affairs and Trade

REVIEW OF 2017 PROGRAM EVALUATIONS

Prepared by the Office of Development Effectiveness (ODE)

Foreword

I am pleased to endorse the Review of Program Evaluations completed in 2017. This Review examines the quality of the independent evaluations conducted by aid program areas in the Department of Foreign Affairs and Trade (DFAT). This is the third such review with previous reviews conducted in 2012 and 2014. The Review of Program Evaluations provides insights into how DFAT's evaluation system is working. It is one of the strengths of the broader aid performance management system.

Program evaluations continue to be a credible source of evidence for the aid program while using modest consultant and financial resources. However, there is room for improvement. While the Review found that around 70% of evaluations were adequate or better quality, there has been a small decline in the overall quality of evaluations compared to the 2012 and 2014 Reviews. In particular, attention is required to ensure that high value investments are evaluated effectively.

A key change since the last review in 2014 has been the introduction of a new aid evaluation policy in 2016. The Review shows that the revised policy has led to an impressive increase in publication rates and management response rates of program evaluations. This is an important foundation for evaluations to be used to their full potential.

The Review makes four recommendations to further improve the quality of program evaluations. The Independent Evaluation Committee (IEC), which I chair, endorses these recommendations. They represent practical, realistic actions which will improve program evaluations. I urge the Office of Development Effectiveness and senior managers in DFAT to act on these recommendations.

Jim Adams,
Chair, Independent Evaluation Committee

Executive summary

In the Department of Foreign Affairs and Trade (DFAT), independent program evaluations are conducted on aid investments managed by country, regional and thematic programs.

The Review of 2017 Program Evaluations assessed and analysed 34 program evaluations which were identified in the DFAT Aid Evaluation Plan and completed in 2017. The Review provided an opportunity to assess the impact of the revised Aid Evaluation Policy (2016) on evaluation practice, quality and use.

The objectives of the Review are:

- to better understand the practices and the quality of independent program evaluations and how these have changed over time by comparing to findings of similar reviews conducted in 2012 and 2014; and
- to provide information to support good quality, independent evaluations across the department.

The key audiences for this Review are Office of Development Effectiveness (ODE), senior managers and program staff commissioning evaluations, and Aid Management and Performance Branch.

The Review was undertaken largely in-house by the Office of Development Effectiveness (ODE). Methods included:

- a desk study to assess the quality of program evaluations against nine criteria, and to identify good practice evaluations and useful lessons on aid and development
- a brief survey of DFAT staff to examine the department's use of 2017 evaluations

Executive summary

There are four main findings from the Review.

Program evaluations use modest financial and consultant resources.

- The median cost of a program evaluation is $89,000.
- Program evaluations cost, on average, 0.86% of investment value.
- The average duration of a program evaluation is 69 days, including 29 fieldwork days.

Program evaluations are a credible source of evidence for the aid program but there is room for improvement.

- 71% of evaluations were assessed as adequate or better quality, however the overall quality of evaluations has declined a small amount when compared to 2012 and 2014 Reviews.
- Quality related to the design elements of evaluations such as scope and methodology have improved over time but quality related to the "core essentials" such as use of program monitoring data and credibility of evidence and analysis has declined since 2012.
- Use and quality of investment monitoring systems was the weakest of the performance criteria, showing a 21% decline in quality compared to 2014.
- The Review also found that a significant number of investments valued above $100 million (three of seven) had inadequate evaluations.

Evaluations that are fit for purpose and actively managed are more likely to be good quality.

- The Review undertook correlation analysis to determine if underlying factors such as duration, cost, team size and composition influenced evaluation quality and found there was no clear association between the majority of these factors and evaluation quality.
- Qualitative analysis showed that the way evaluations are planned and managed is the largest determining factor of evaluation quality.

Executive summary

Findings from program evaluations are being used to improve implementation and inform future aid designs.

- Publication and management response rates have strongly increased since the introduction of the revised Aid Evaluation Policy in 2016.
- Responses from a qualitative survey indicated that a large number of recommendations from 2017 evaluations are being implemented to improve ongoing programs and inform future aid designs. Major areas where recommendations were being implemented were in strengthening gender and social inclusion strategies and program monitoring and evaluation systems.

The findings show that some of the key objectives of the revised Aid Evaluation Policy have been achieved.

- The revised aid evaluation policy has been successful in addressing poor publication and management response rates, a key finding in the 2014 Review.
- The revised aid evaluation policy has laid the foundation for evaluations to be better used to their full potential.

The objective of the evaluation policy to improve the quality of program evaluations has not yet been achieved. It is difficult to identify reasons for this from data collected and analysed in this Review. Although the overall quality of evaluations has declined, practice related to the design of evaluations has improved since 2012. Design of evaluations was found to be weak in the 2012 Review and recommendations were put in place to ensure better planning of evaluations.

It is clear that further action is required to improve the quality of program evaluations, particularly for larger value investments. Action should be focused on improving the “core essentials” of evaluations, including through stronger oversight and management by DFAT. This needs to be complemented by strengthening investment monitoring systems to ensure robust and credible data is available to measure program performance.

Executive summary

To address these findings, it is recommended that ODE:

1) engage more closely with Divisions on their consideration of evaluations to be nominated for the DFAT annual aid evaluation plan
2) identify ODE contact officers for each relevant Division to provide guidance on evaluation requirements and support evaluation capability
3) review terms of reference, evaluation plans and draft reports for investments valued at, or greater than $50 million
4) liaise with the Contracting and Aid Management Division to consider options for strengthening investment monitoring systems to deliver more robust and credible data.

CONTENTS

BACKGROUND AND METHODS

The Review of Program Evaluations aims to improve our understanding of the practices relating to, and quality of, program evaluations

In the Department of Foreign Affairs and Trade (DFAT), independent evaluations are undertaken at two levels:

- strategic evaluations are produced by the Office of Development Effectiveness (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. Each program undertakes an annual process to identify a minimum number of evaluations to address the highest priority information needs. These largely focus on an individual aid investment

This Review focuses on independent program evaluations. It has three objectives:

- to better understand the practices related to, and the quality of, independent program evaluations and how these have changed over time by comparing to findings of similar reviews conducted for 2012 and 2014 evaluations
- to provide information to support good quality, independent evaluations across the department
- to promote better use of evaluations across the department and the aid community by facilitating opportunities for learning.

The Review was conducted in two phases.

- Phase 1: Quality review of all independent program evaluations completed and published in 2017 and a brief survey of the use of 2017 evaluations
- Phase 2: Synthesis and dissemination of lessons from 2017 program evaluations.

This report covers the quality and use of evaluations (Phase 1). Separate briefs will be produced to report on the synthesis of lessons from 2017 evaluations (Phase 2).

The Review provides an opportunity to assess the impact of the revised Aid Evaluation Policy (2016) on evaluation practice, quality and use

Since 2012, the aid evaluation policy has shifted from an approach which required all investments over a particular value threshold or risk profile to be evaluated using specific quality criteria to a demand-driven approach where program areas have the flexibility to determine what evaluations they will conduct to meet their information needs.

Requirements under the revised Aid Evaluation Policy (2016) include: (1) each country/regional/thematic program is given a minimum number of evaluations to be conducted each year or every few years, with larger programs expected to undertake more evaluations; and (2) each year Division Heads nominate evaluations they will undertake. ODE then compiles DFAT's Annual Evaluation Plan which is reviewed and approved by the Secretary and shared with the Minister for Foreign Affairs.

The main audiences for the Review are ODE, and managers and staff commissioning evaluations

ODE will use the Review's findings on the quality of program evaluations to improve DFAT's Aid Evaluation Policy implementation.

Senior managers and staff commissioning evaluations in country and regional programs will also use the Review's findings to manage high quality program evaluations.

Secondary audiences include DFAT's Aid Management and Performance Branch (MPB), which oversees DFAT's investment quality reporting system (of which program evaluations are part).

This Review builds on previous Reviews of Program Evaluations conducted by ODE

ODE has completed two previous Reviews of Program Evaluations:

- the first Review used contracted consultants to review all 87 independent program evaluations completed in 2012.
- the second Review was conducted by the ODE and reviewed 35 program evaluations conducted in 2014. This was a purposeful sample from the 77 program evaluations completed in 2014.

This Review examined all 37 program evaluations identified in the Aid Evaluation Plan and completed in 2017.

- This allowed the ODE to draw conclusions that apply to all program evaluations completed in 2017 and to assess whether 2017 program evaluations included adequate coverage of the aid program in terms of sectors, geographic focus and funding.
- The smaller number of evaluations in 2017 was an intended effect of the revised evaluation policy, which is testing whether fewer, demand driven evaluations with greater senior management oversight results in better quality evaluations.

This Review was conducted largely by ODE staff.

- This was to ensure the findings would be relevant to DFAT; any proposed follow-up actions would be appropriate and feasible; and ODE staff would gain a strong understanding of current evaluation practice in the department.
- A contracted consultant, who was part of the Review team, assisted with assessing the quality of 2017 evaluations (Phase 1) and led the work on the synthesis of lessons (Phase 2). The synthesis reports will be published separately to this report.

The Review examines seven key evaluation questions

The main method was a desk review of evaluation Terms of Reference (TORs), plans and reports. We also conducted a qualitative survey on the use of 2017 evaluations.

— The evaluation questions, and the methods used to address them, are summarised in the table below.

Evaluation question	Data collection and analysis methods
Priority questions	
1. What are the characteristics and quality of program evaluations? How have these changed since 2012 and 2014?	• Basic characteristics of 2017 evaluations (e.g. size of investment, number of days, team size etc) collected from Review Proforma and DFAT's aid management system (Aidworks). • Evaluations rated against quality criteria. • Analysis of each criteria to establish areas where evaluation quality is high or low. • Evaluation characteristics and quality compared to the 2012 and 2014 Reviews. • Establish a measure of overall evaluation quality.
2. What factors contribute to the quality of program evaluations?	• Correlation analysis to examine relationships between evaluation quality and possible factors contributing to evaluation quality collected under Q1. • Qualitative analysis of narrative accompanying scores on Review Proforma.
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 criteria "credibility of evidence and analysis", which was the proxy indicator for overall quality
Other evaluation questions	
4. How are the findings from program evaluations used in the department?	• Management responses assessed to examine the proportion of recommendations accepted. • Short survey to relevant staff to identify (1) how evaluation recommendations are being used to influence policy and program development; and (2) constraints to implementing evaluation recommendations.
5. Which evaluations can be nominated for a Secretary's award for evaluation excellence.	• A small number of good practice evaluations will be identified and recommended for a Secretary's award for evaluation excellence.
6. Based on the findings of this Review, what are the implications for the Department's evaluation policy?	• Data and analysis from Q1, Q2, Q3 and Q4 will be collated and analysed to identify where DFAT's evaluation policies and practices could be adjusted.
7. What can be learned 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.	• The Review Team identified sections in the evaluation reports that provide learning on policy influence, aid capability and gender equality. The synthesis of learning was prepared by a consultant and will be reported in a series of separate briefs.

To determine evaluation quality, each program evaluation was assessed against nine quality criteria

The criteria (summarised below) are based on DFAT's Monitoring and Evaluation Standards.

- For each criterion a program evaluation was given a score between 1 (very poor quality) and 6 (very high quality).
- The same criteria were used as in previous Reviews in order to compare changes over time.

Quality criteria	Summary description
1) Executive Summary	The executive summary provides all the necessary information (including on gender issues) to enable primary users to make good quality decisions.
2) Purpose of evaluation	The purpose of the evaluation is provided, including the overall purpose and primary users of the information.
3) Scope of evaluation	The scope of the evaluation matches the evaluation resources. Data collection methods are defined and take into account the needs of groups such as women.
4) Appropriateness of methodology and use of sources	Justification is provided for the data collection and analysis techniques chosen. Consideration is given to analysis of sex-disaggregated data. Triangulation is sufficient and the sampling strategy is appropriate. Limitations to the methods and ethical issues are described and addressed.
5) Adequacy and use of M&E	The adequacy of M&E data and systems are described. Where good quality data is available and relevant to evaluation questions, the evaluation makes use of it.
6) Context of the investment	The context of the investment (including relevant gender issues) is described and its influence on performance is assessed.
7) Evaluation questions	The evaluation identifies appropriate evaluation questions and then answers them. An appropriate balance is made between operational and strategic issues.
8) Credibility of evidence and analysis	Findings flow logically from the data, showing a clear line of evidence. Conclusions, recommendations and lessons are substantiated by findings and analysis. Findings relevant to specific groups such as women are included.
9) Recommendations	Conclusions, recommendations and lessons are clear, relevant, targeted and actionable.

There were five main limitations to the Review

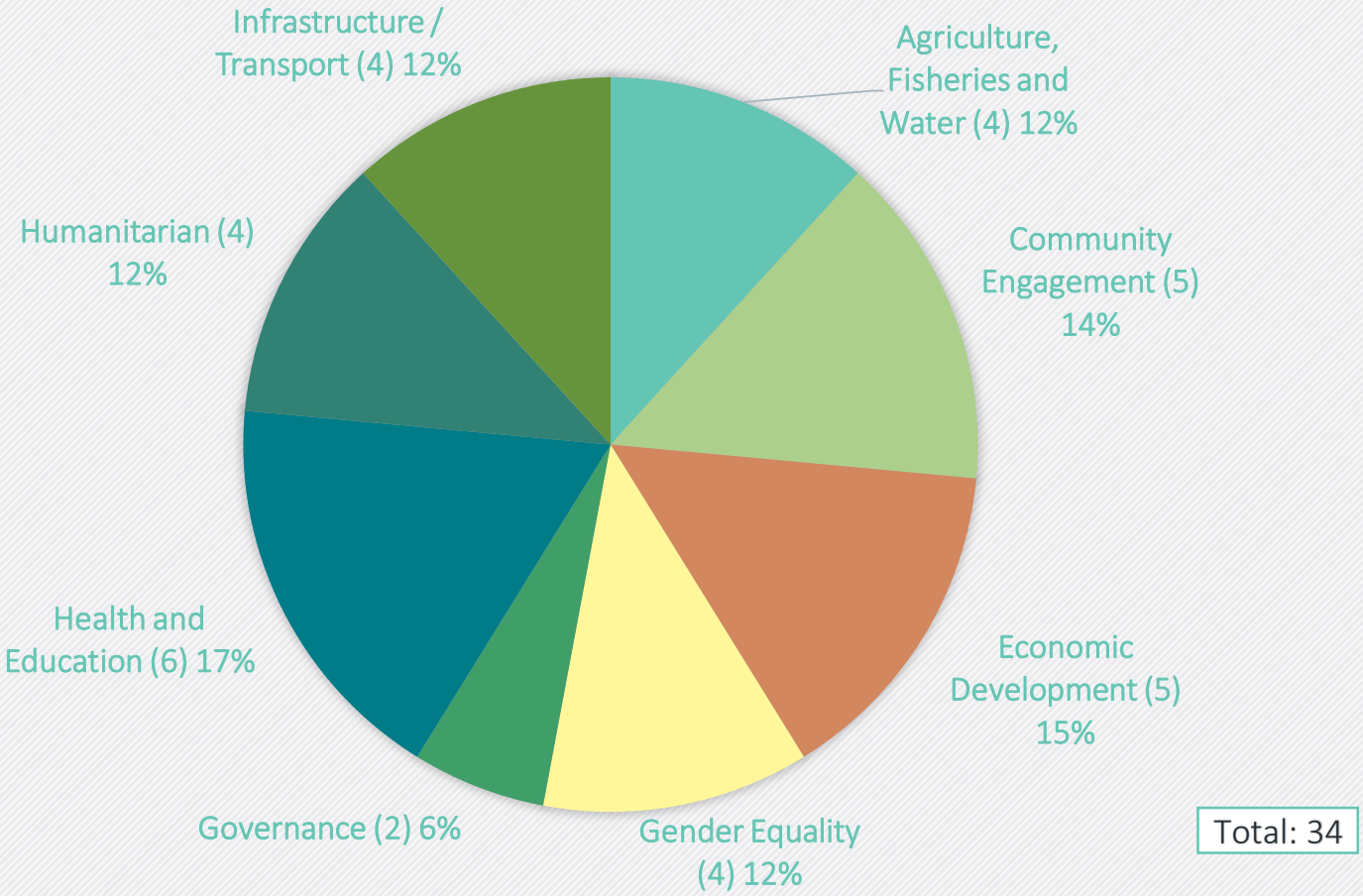
These limitations, and how they have been addressed, are summarised in the table below.

Limitation	Management strategy
ODE is assessing the evaluation policy, guidance and support which it oversees and will need to respond to the Review's findings. It will be difficult for ODE to draft recommendations as well as a management response to the recommendations.	The Independent Evaluation Committee (IEC) will oversee the Review to help ensure any self-assessment conducted by ODE is defensible. The IEC will also assess the quality and feasibility of the recommendations. If the IEC supports the recommendations, ODE will implement them. No formal management response will be completed.
Review team members need to assess program evaluations in a consistent manner.	A Review Handbook outlined in detail how assessments should be undertaken. A number of moderation meetings were held during the assessment process to ensure team members were assessing evaluations consistently.
The number of evaluations, although a census or full population of 2017 evaluations, was small in size (34), making it difficult to determine statistical significance. The small population size means that small changes in numbers can result in relatively larger changes in percentages.	Findings and conclusions need to be interpreted with this limitation in mind. Quantitative findings are presented in both numbers and percentages.
Comparing findings with 2014 Review posed limitations in that the 2014 Review used a purposeful sample whereas the 2017 Review used a census or full population of evaluations completed in that year. The 2012 Review, although a larger number of evaluations (87), was also a census making results more comparable.	Comparative analysis with the 2014 Review needs to be treated with caution and this is highlighted throughout the report.
Whilst assessing all 37 program evaluations completed in 2017, three reviews were found not to meet DFAT evaluation requirements, i.e. they were not an independent or a systematic and in-depth assessment of a program.	The three evaluations were excluded from the analysis and reporting. One was an internal review by an implementing partner and two were annual reviews to trigger performance linked funding. This means we would not have been comparing "like with like" when assessing the characteristics of program evaluations, and we therefore excluded them from the quality analysis. This reduced the number of evaluations in this Review to 34.

CHARACTERISTICS OF 2017 EVALUATIONS

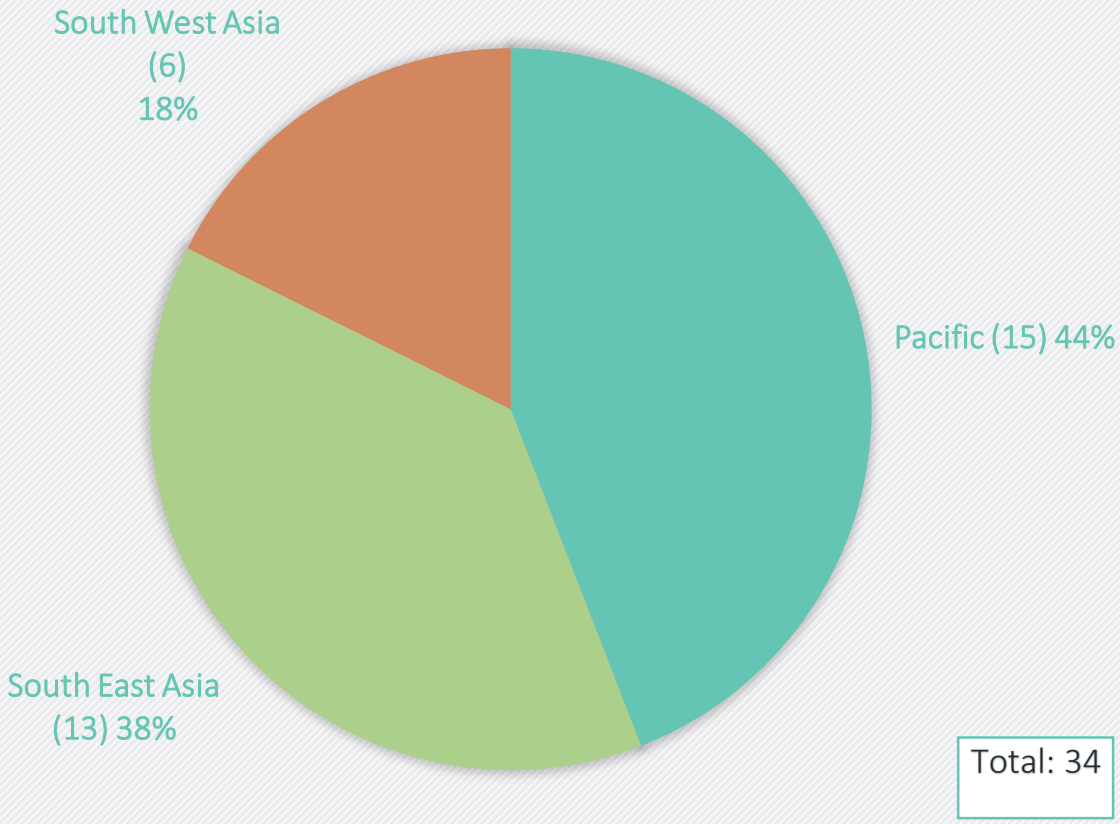
2017 EVALUATIONS BY AID SECTOR

The range of aid program sectors covered by the evaluations aligns with DFAT's key sectoral priorities identified in *Australian Aid: promoting prosperity, reducing poverty, enhancing stability*



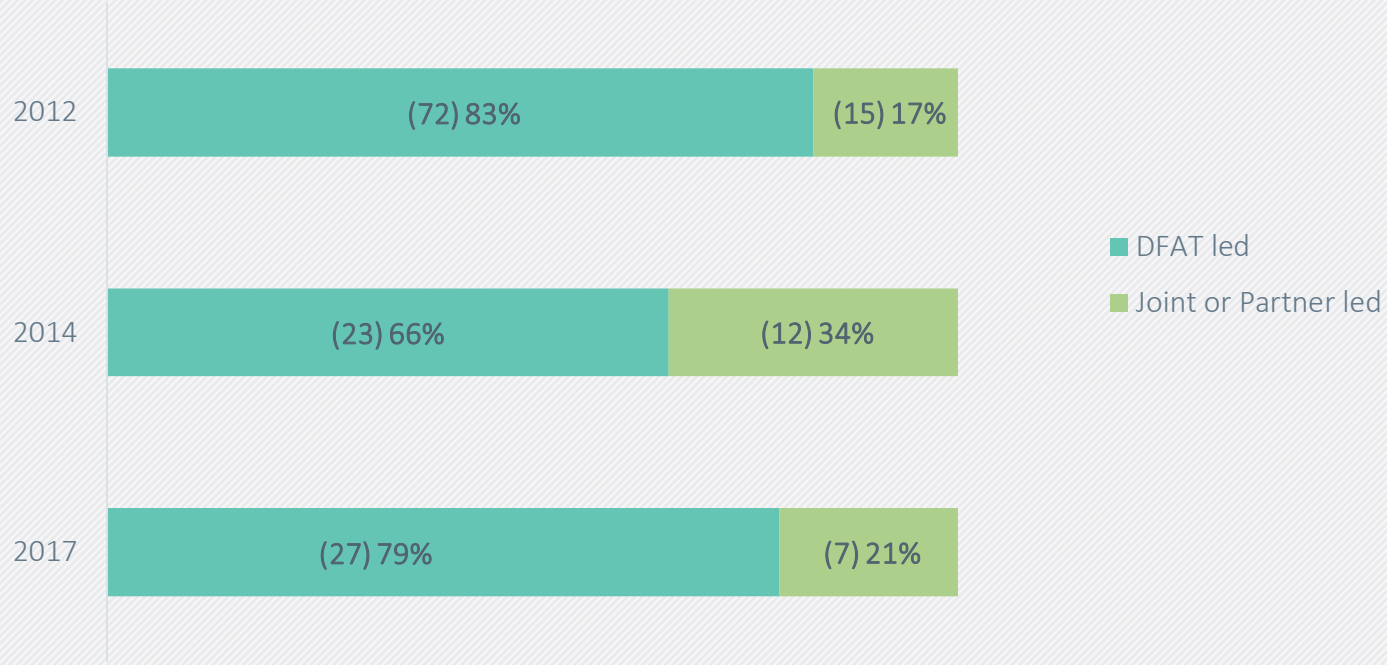
2017 EVALUATIONS BY GEOGRAPHIC REGION

The coverage of evaluations by region reflects geographic priorities identified in *Australian Aid: promoting prosperity, reducing poverty, enhancing stability*.



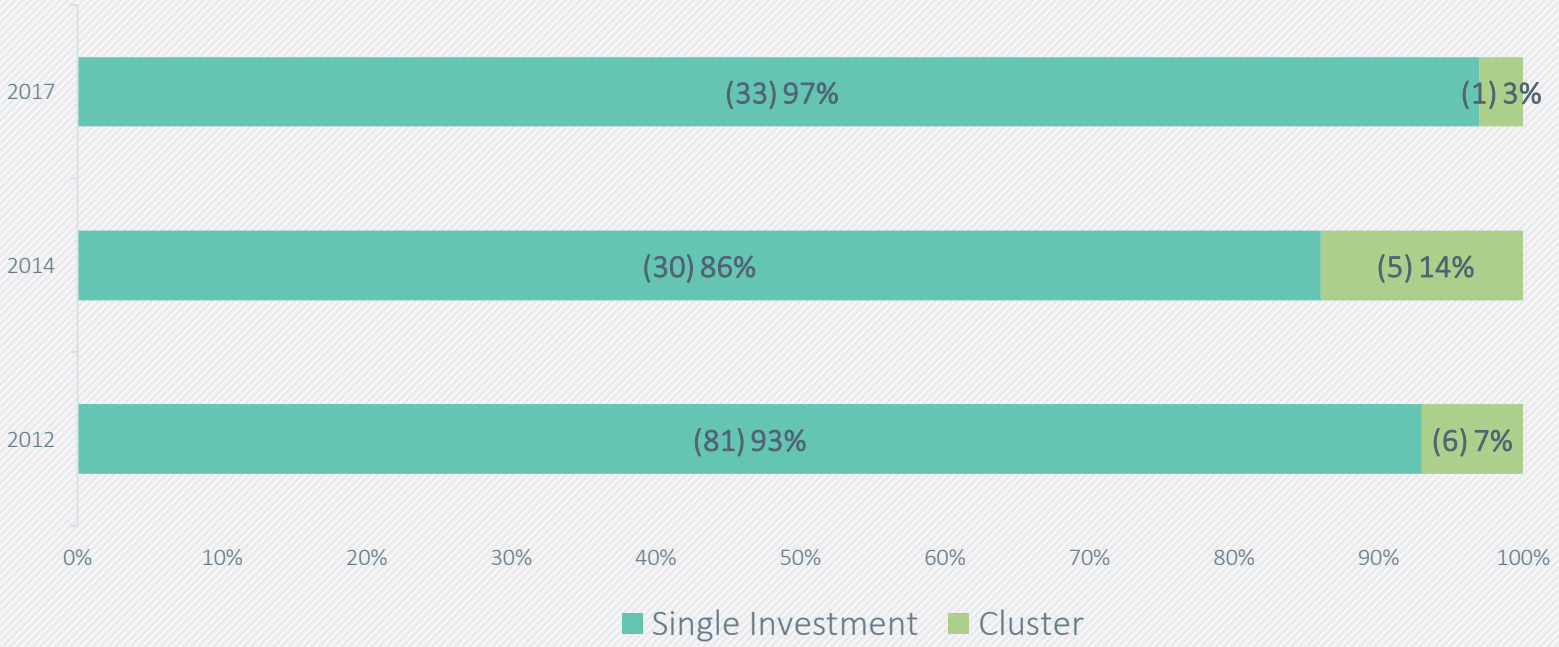
THE MAJORITY OF EVALUATIONS ARE DFAT-LED

The proportion of evaluations commissioned by one of DFAT's partners (for example, another donor), or conducted jointly by DFAT and a development partner, has fluctuated slightly since 2012.



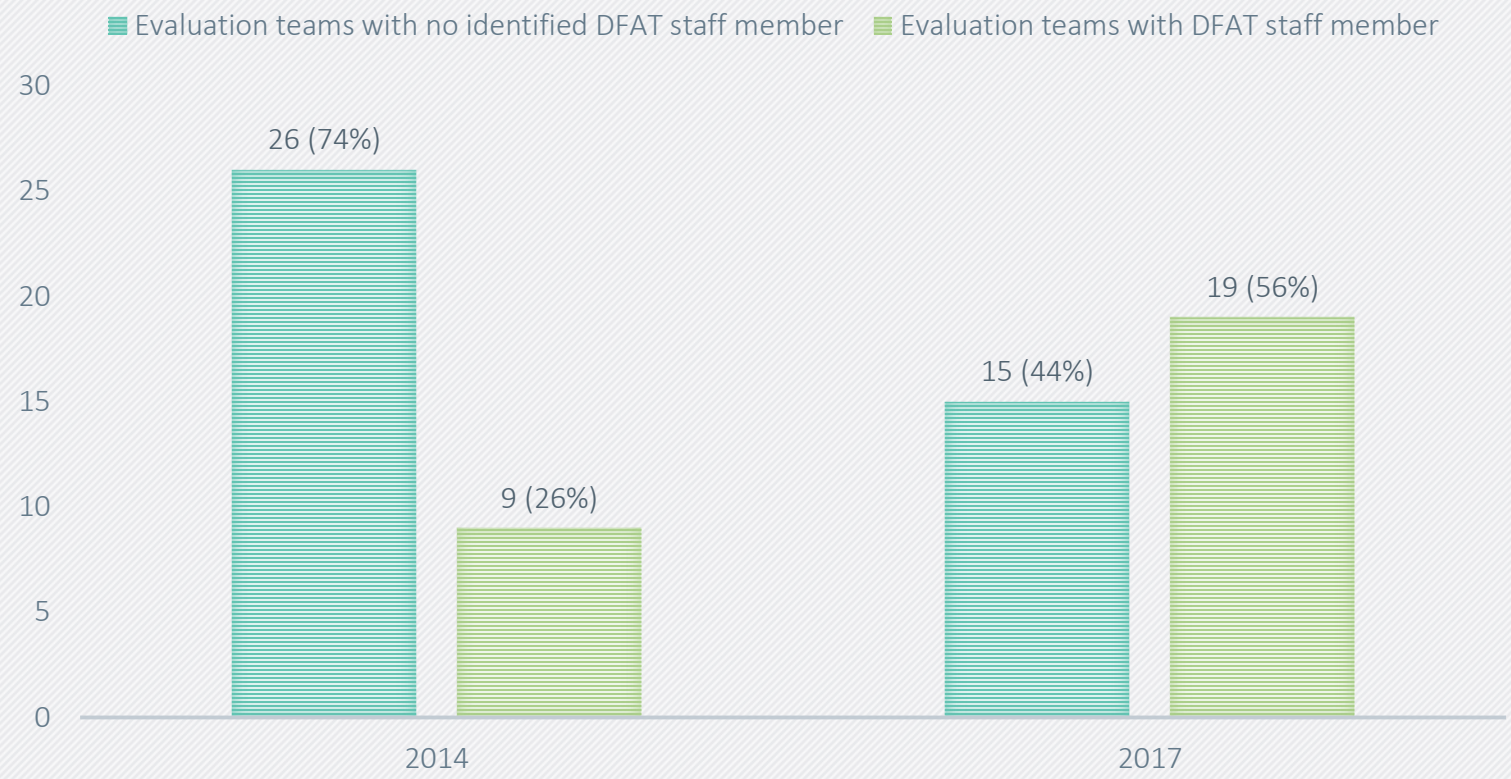
THE VAST MAJORITY OF EVALUATIONS ARE SINGLE INVESTMENT EVALUATIONS

The proportion of cluster investment evaluations has fluctuated by a small amount since 2012.



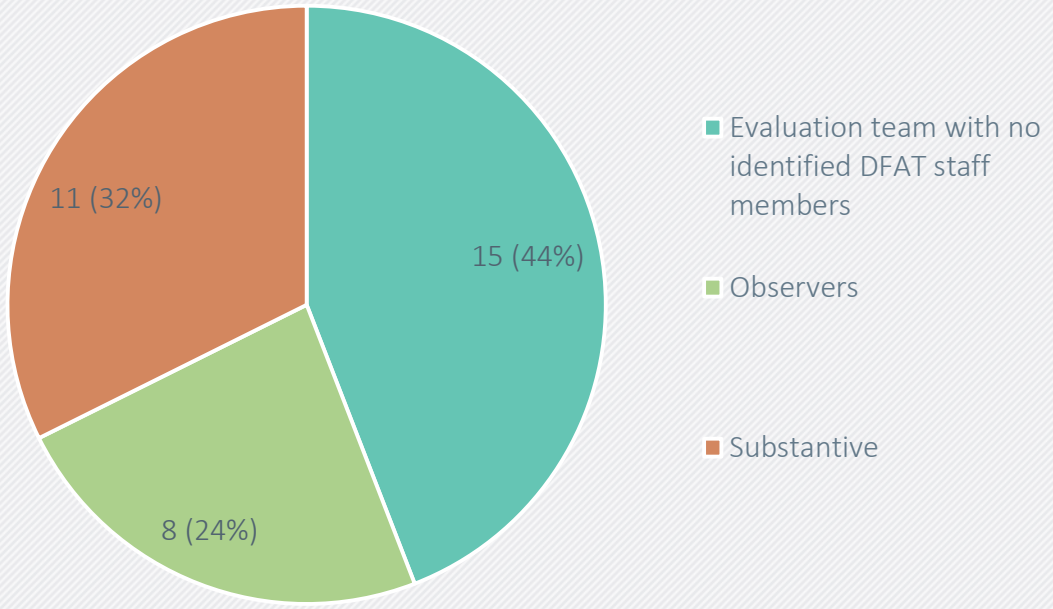
EVALUATION TEAMS WITH DFAT STAFF HAVE INCREASED

The 2014 Review recommended DFAT increase staff involvement in evaluations to strengthen the likelihood that evaluations are relevant and recommendations are appropriate. The graph shows that in 2017 there was a higher proportion of evaluations with DFAT staff on the evaluation team compared to 2014. Evaluation teams with DFAT staff were slightly more likely to be rated adequate or better quality.



ROLE OF DFAT STAFF MEMBER ON EVALUATION TEAM

In 2017 DFAT staff members on evaluation teams were more likely to play a substantive role (11) rather than an observer role (nine). A substantive role included being an active team member with a significant role in the implementation of the evaluation such as in data collection or report writing.



KEY FINDINGS

1. Program evaluations use modest financial and consultant resources

2. Most evaluations are a credible source of evidence for the aid program but there is room for improvement

3. Evaluations that are fit for purpose and actively managed by DFAT are more likely to be good quality

4. The findings from program evaluations are being used to improve implementation and inform future aid designs

FINDING 1: Program evaluations use modest financial and consultant resources

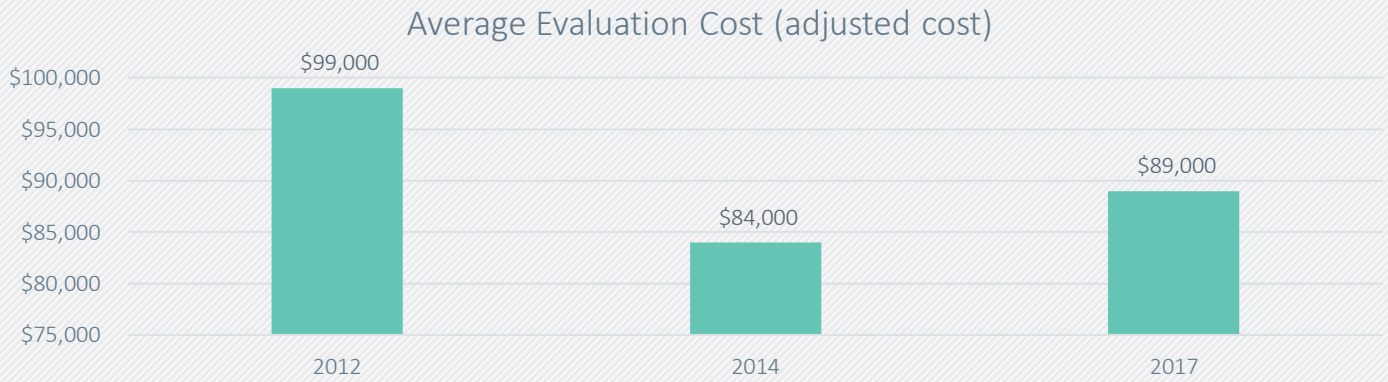
FINDING 1: Program evaluations use modest financial and consultant resources

Cost of program evaluations was modest.

The average cost of the evaluations was $89,000.

- This represents consultant costs only and not the cost of DFAT staff time.
- This represents a slight increase since 2014. When adjusted for inflation, 2014 spending was $84,000.
- Compared to 2012, spending declined a little but remains within a similar range. When adjusted for inflation, 2012 spending equalled $99,000.

Program evaluations cost between 0.02% and 2.5% of investment value. The median evaluation cost as a proportion of investment value was 0.86%.



FINDING 1: Program evaluations use modest financial and consultant resources

Average person working days* and field work days remain modest.

- On average, the total number of **working days** committed to a program evaluation by all consultants was **69**. This compares with 72 in 2014.
- It is important to note that consultant contracting information was not accessible for a large number of evaluations so it was inferred from the information available, consistent with the approach in 2014.
- On average, the total number of **fieldwork days** committed to a program evaluation by all contracted team members was **29**. This compares with 32.5 in 2014.

* Average working days and fieldwork days only included consultants time so we could compare to 2014 data. As noted previously, DFAT staff involvement in evaluation teams was greater in 2017, however this has not been quantified. The number of staff input days is more difficult to infer than consultants due to their more complex role in scoping and managing the evaluation.

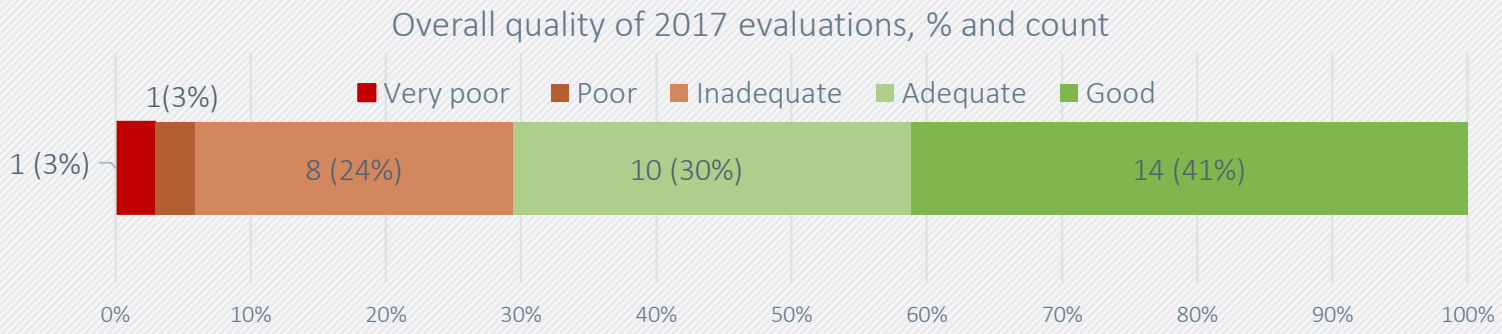
FINDING 2: Most evaluations are a credible source of evidence for the aid program but there is room for improvement

FINDING 2: Most evaluations are a credible source of evidence for the aid program but there is room for improvement

Credibility of evidence and analysis was used as the proxy measure for assessing **overall quality of evaluations**. We chose this criteria because:

- correlation analysis demonstrated there was a strong positive relationship between this and the other eight criteria
- this criteria focuses on the sound evidence base of an evaluation. Using a common sense test, this is a good indicator for overall evaluation quality
- the 2012 and 2014 Reviews used this criteria to represent overall quality of each program evaluation. Using it again allows for easier comparison between the current and previous Reviews.

71% of 2017 program evaluations were assessed as of adequate or better quality (a score of 4 or more on the 6 point rating scale) for credibility of evidence and analysis.

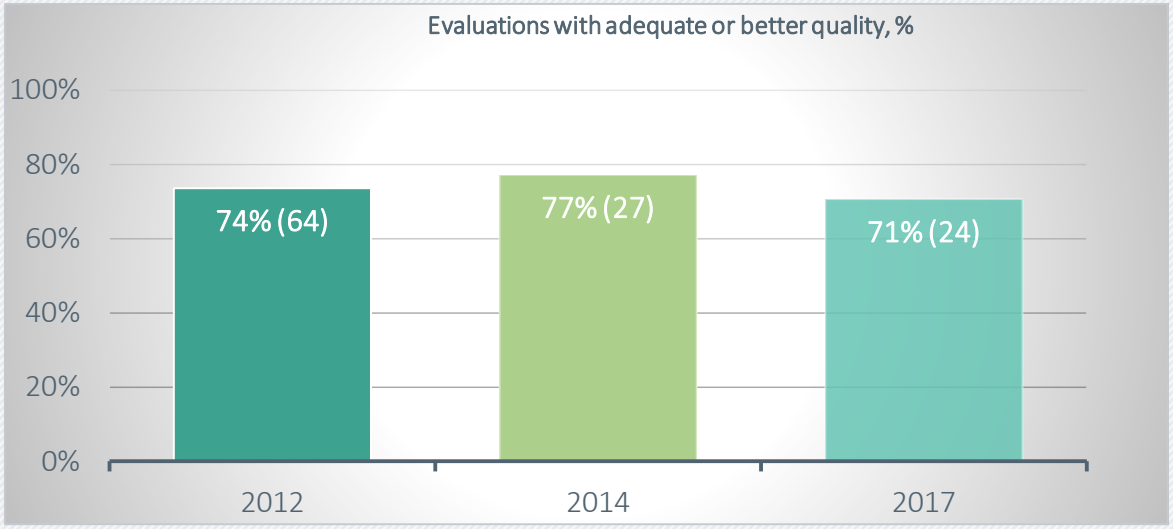


FINDING 2: Most evaluations are a credible source of evidence for the aid program but there is room for improvement,

The 2012 and 2014 Reviews concluded, using the same proxy indicator for overall quality, that program evaluations were satisfactory and a credible source of evidence for the aid program.

However, the **overall quality of evaluations** has declined a small amount from 2014 and 2012 ratings.

- Although the decline is small, ODE expected that the quality of program evaluations would improve under the new Aid Evaluation Policy, which introduced a demand-driven approach and more senior oversight of evaluations. The review is not able to establish the reason for the decline from data collected, as this falls beyond the Review's scope.
- The small size (34) and the difference in approaches between 2014 and 2017 (sample versus population) make it difficult to judge whether this difference is statistically significant.

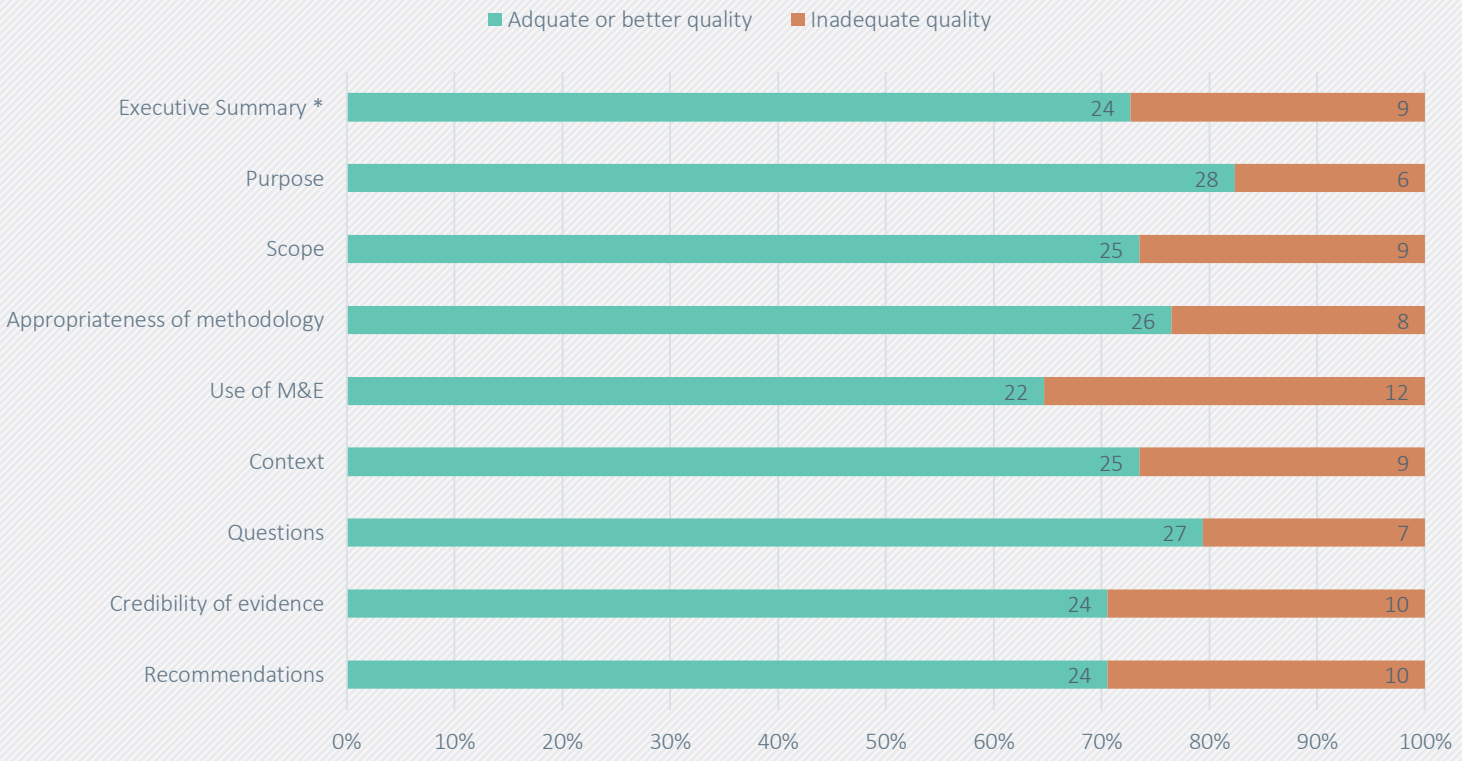


FINDING 2: Most evaluations are a credible source of evidence for the aid program but there is room for improvement

For **eight of the nine** criteria, at least 70% of evaluations were assessed as adequate or better quality (4 – 6 on the rating scale).

FINDING 2: Most evaluations are a credible source of evidence for the aid program but there is room for improvement

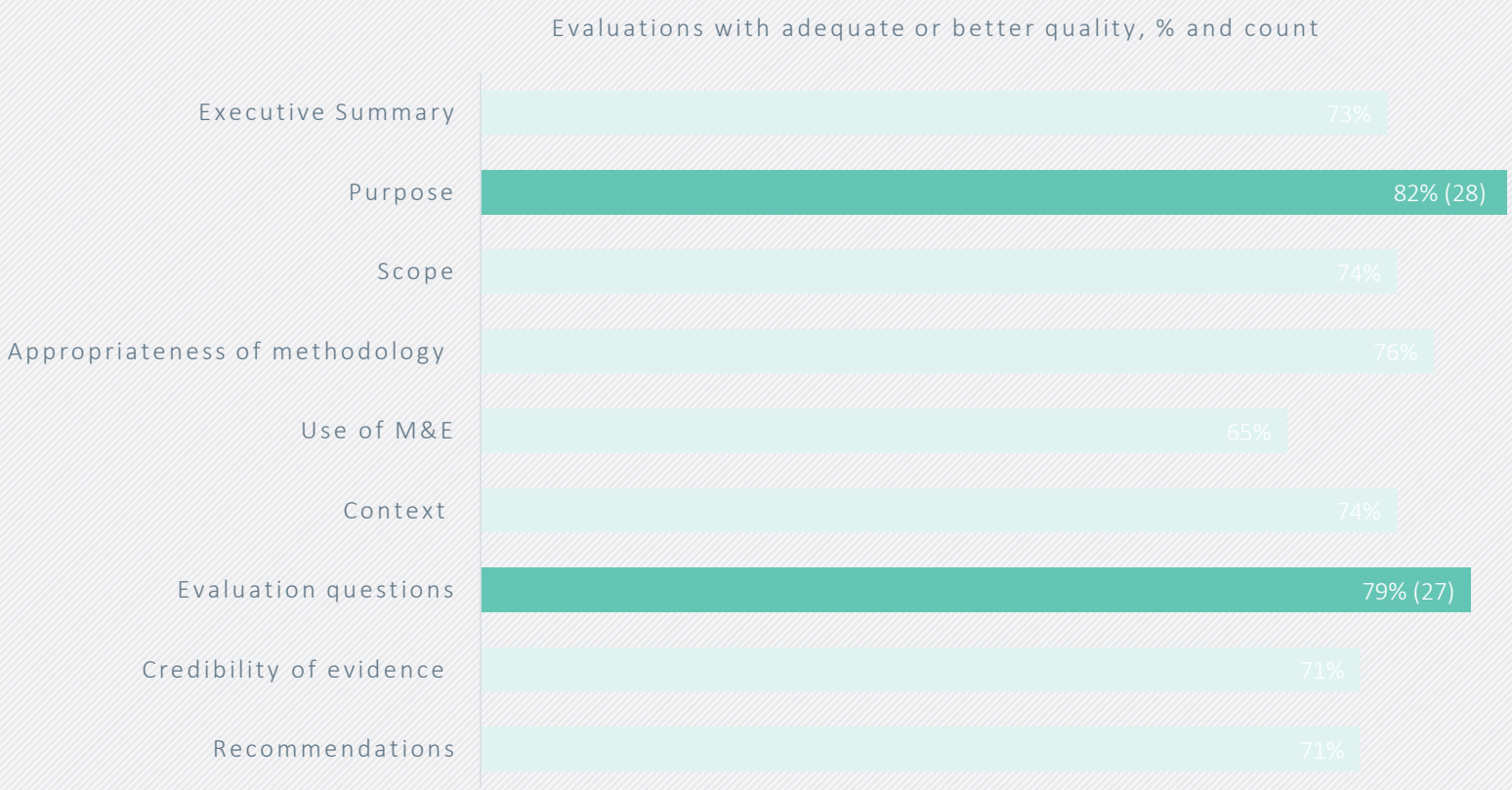
The graph shows the number of evaluations that were rated adequate or better quality and inadequate for each quality criteria.



* Only 33 of the program evaluations had Executive Summaries

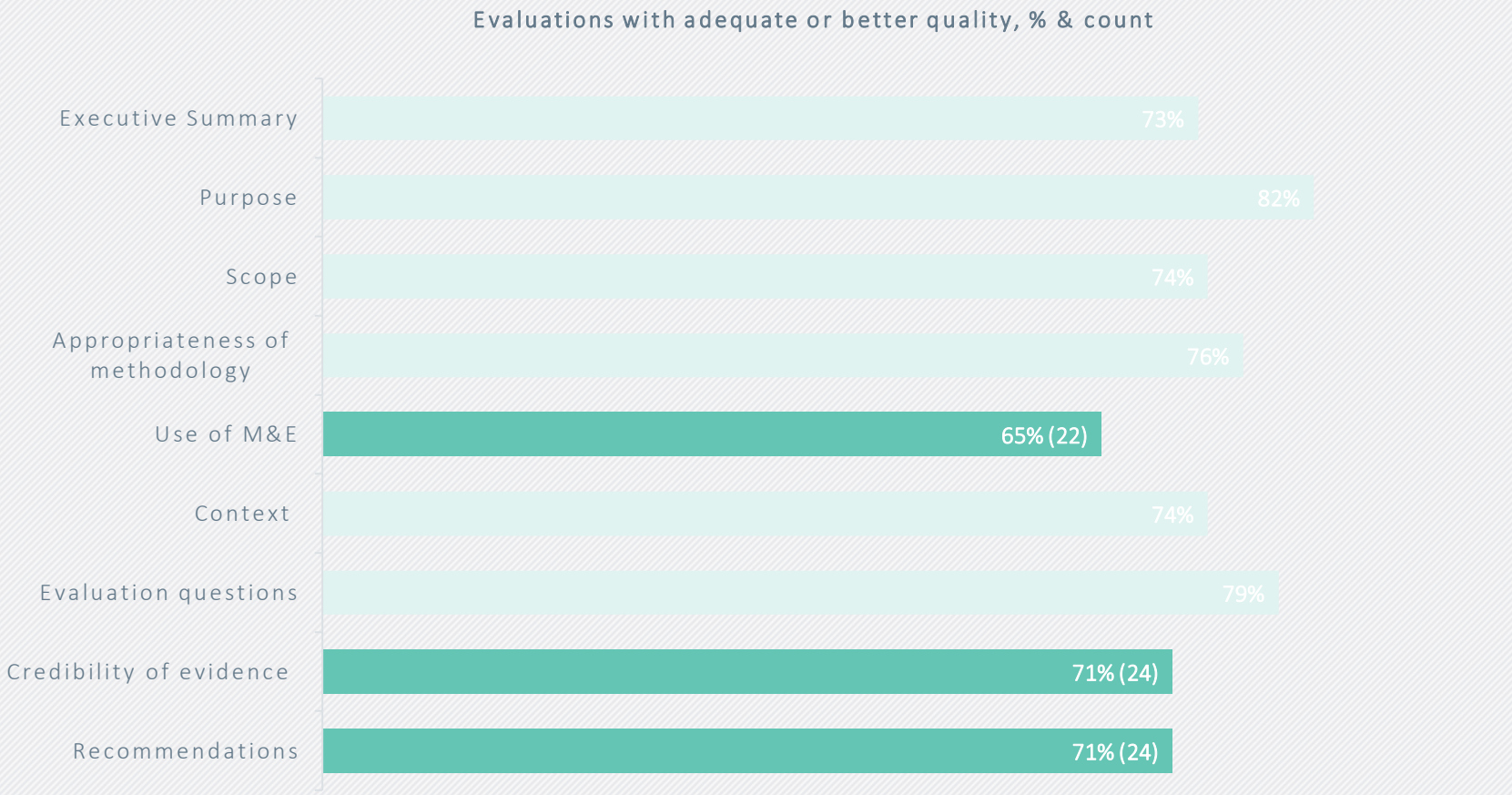
FINDING 2: Most evaluations are a credible source of evidence for the aid program but there is room for improvement

There were strong results for two criteria: for **purpose** and e**valuation questions:** 82% and 79% of evaluations were assessed as adequate or better quality.



FINDING 2: Most evaluations are a credible source of evidence for the aid program but there is room for improvement

The criteria with the lowest percentage of evaluations with adequate or better quality were **use of M&E** (65%) and **credibility of evidence** and **recommendations** (71%).

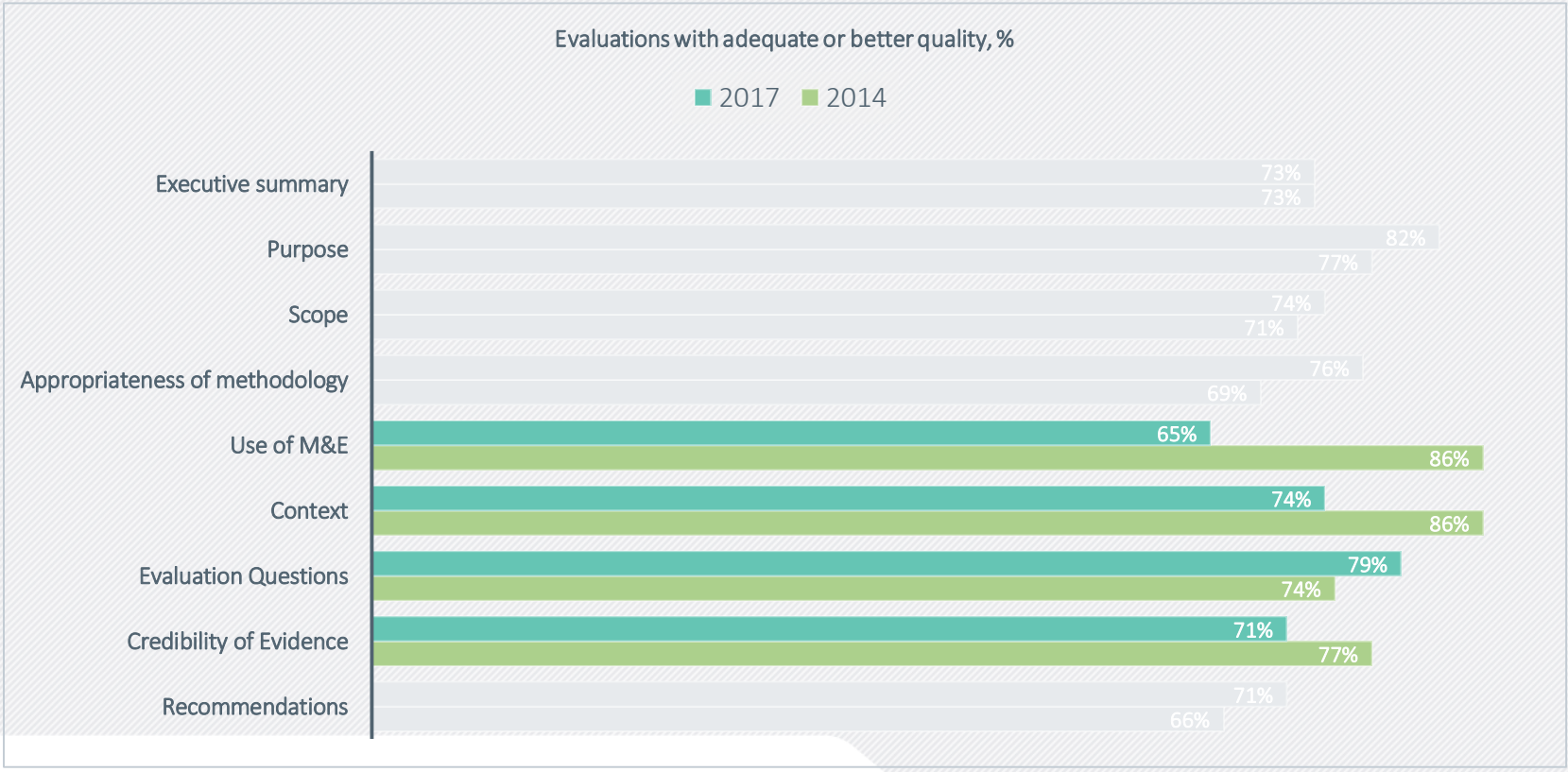


FINDING 2: Most evaluations are a credible source of evidence for the aid program but there is room for improvement

- Compared to 2014, **a greater proportion of evaluations were assessed as adequate or better quality** against criteria that relate to the **design** of evaluations.
- These include purpose, scope, appropriateness of methodology and evaluation questions.

FINDING 2: Most evaluations are a credible source of evidence for the aid program but there is room for improvement

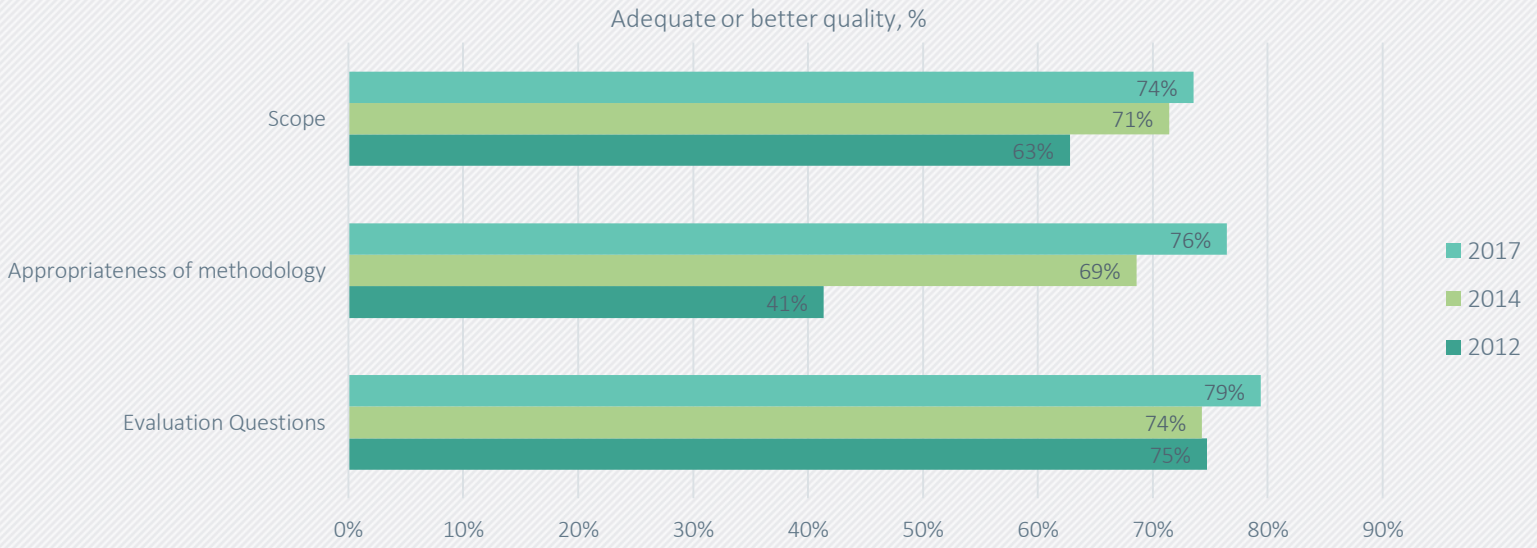
- Compared to 2014, **a lower proportion of evaluations were assessed as adequate or better quality** against criteria that relate to the **"core essential" elements of robust evaluations**: use of M&E, analysis of context and credibility of evidence.
- The **21% decline for adequacy and use of M&E** since 2014 is noteworthy. This is consistent with findings from the 2017 AQC spot check and the recent ODE Review of Investment Monitoring.



FINDING 2: Most evaluations are a credible source of evidence for the aid program but there is room for improvement

We also compared findings to those of the first quality review of evaluations conducted in 2012 to track changes over time. The graph below shows evaluation components that have improved compared to 2012 and 2014.

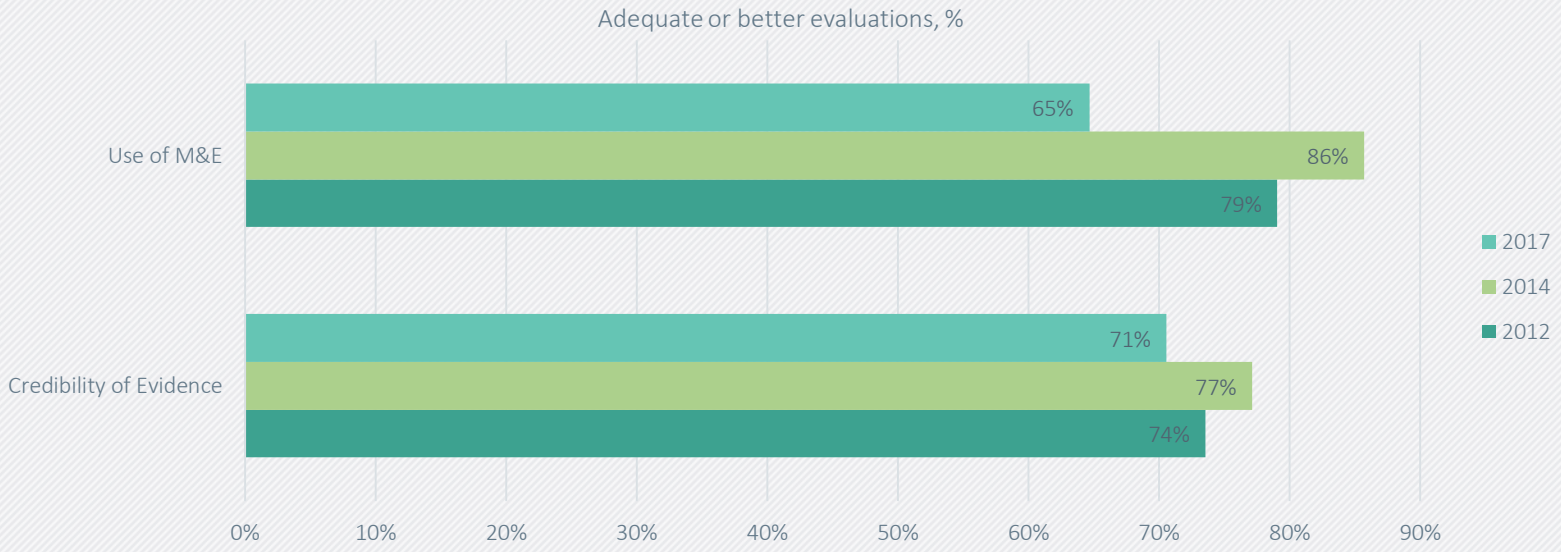
- **Scope** and **methodology** are the only two criteria that have **improved** consistently since 2012. Methodology has shown the greatest improvement of 35% since 2012.
- **Evaluation questions** have improved compared to both 2012 and 2014.
- The above criteria relate to the **design aspect of evaluations**. The 2012 review found that the design components of evaluations were weakest compared to other elements and recommended that DFAT focus greater support and quality assurance efforts at an early stage of an evaluation.
- However, caution should be used in interpreting the data due to the small population size and the difference in approaches between 2014 and 2017.



FINDING 2: Most evaluations are a credible source of evidence for the aid program but there is room for improvement

Compared to 2012 and 2014, the percentage of evaluations assessed as adequate or better quality for **use of M&E** and **credibility of evidence and analysis** are **lower in 2017.**

- The decline for credibility of evidence and analysis does not represent a strong change - only three and six percentage points compared to 2012 and 2014.
- Again, caution should be taken in interpreting the data due to the small population size and the difference in approaches between 2014 and 2017.
- As highlighted earlier, the 21% decline in quality and use of M&E between 2014 and 2017 Reviews is particularly noteworthy. This report identifies common themes behind inadequate ratings (see next page).



FINDING 2: Most evaluations are a credible source of evidence for the aid program but there is room for improvement

The quality of program evaluations could be improved by strengthening the use and adequacy of investment monitoring systems.

Content analysis of the 2017 evaluations showed the following common themes and issues in relation to use and adequacy of investment M&E data:

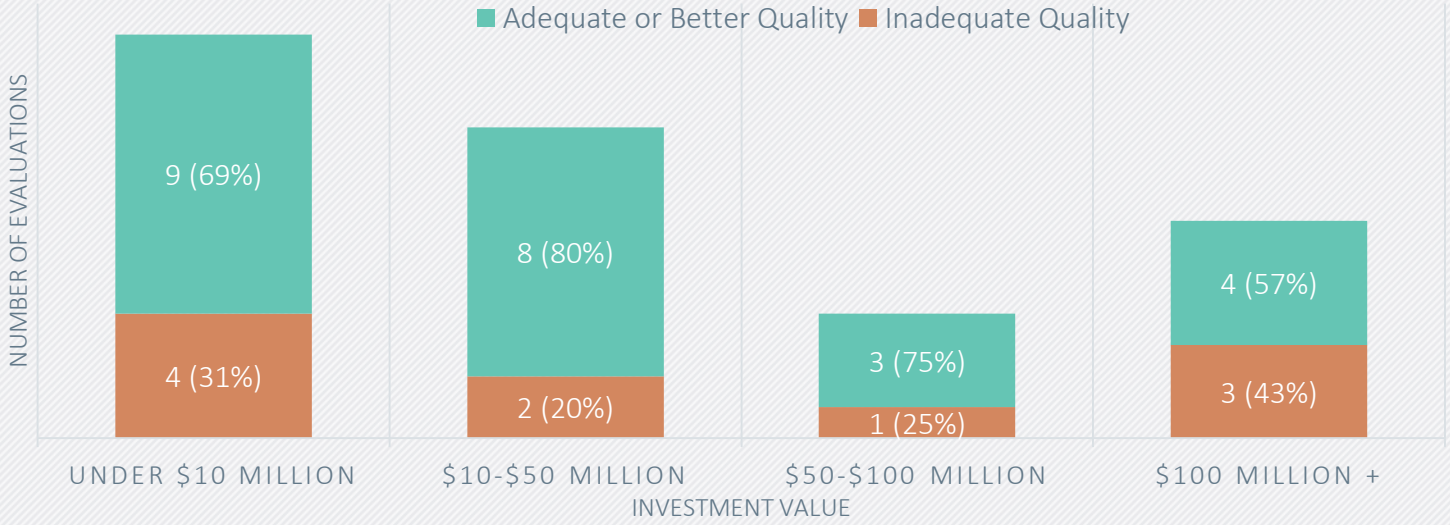
- effective monitoring systems were in place for assessing inputs and outputs (numbers of people trained, amount of seeds distributed) but there were insufficient data for effectively measuring outcomes or impacts of programs
- inadequate methods for monitoring systemic or whole of program changes rather than changes at individual component levels
- M&E systems which were too complex or impenetrable, making their use by DFAT or the partner government difficult
- inadequate reporting from multilateral organisations including multilateral banks and UN organisations
- an over-emphasis on accountability and under-use of monitoring for learning, advocacy, management and planning
- the importance of monitoring systems being strengthened within the partner government systems rather than DFAT developing standalone, parallel systems.

Strengthening investment monitoring systems will ensure that robust and credible data is available to measure program performance. The above themes and issues should be considered when implementing recommendations related to improving investment monitoring systems.

FINDING 2: Most evaluations are a credible source of evidence for the aid program but there is room for improvement

EVALUATION QUALITY AND TOTAL INVESTMENT VALUE

- Although there is no real pattern between evaluation quality and total investment value, the graph shows **a higher proportion of inadequate evaluations in the >$100 million category**. This is a similar finding to 2012 and 2014. A possible explanation for this may be that higher value investments are more complex and multidimensional and therefore more difficult to evaluate. One of these investments was being implemented across a number of countries.
- Although the number of inadequate quality evaluations in the >$100 million category is low (three of seven), the implications of this finding is worthy of further analysis given the large value and complex nature of these investments.



FINDING 2: Most evaluations are a credible source of evidence for the aid program but there is room for improvement

Further attention is required to ensure that high value investments are evaluated effectively.

Drawing from commentary to justify "quality of evidence and analysis" ratings (overall evaluation quality proxy), key issues identified for inadequate quality evaluations of investments valued more than $100 million included the following:

- evaluation was largely descriptive and did not draw data from a range of sources, including M&E data, to present a coherent and convincing position of what was and was not working well and changes required
- ambiguous or insufficient evidence and analysis to support findings, conclusions and recommendations
- too narrowly focused on measuring outputs at the expense of higher level objectives or limited analysis of attribution or contribution of findings at outcome or impact levels
- insufficient analysis of the influence of context-specific needs and challenges in a multi-country program.

These findings suggest a need for greater planning, management oversight and quality assurance processes for evaluations of large value investments.

FINDING 3: Evaluations that are fit for purpose and actively managed by DFAT are more likely to be good quality

FINDING 3: Evaluations that are fit for purpose and actively managed by DFAT are more likely to be good quality

The Review considered a range of underlying factors influencing evaluation quality, including: team composition, team size, number of evaluation questions, evaluation duration, number of field days and commissioning agency. The Review undertook correlation analysis to understand the relationship between these factors and the quality of evaluations.

Overall, the Review found no clear association between the majority of these factors and evaluation quality.

- The main characteristic associated with evaluation quality was commissioning agency. Evaluations that were commissioned by DFAT were more typically associated with evaluations rated adequate or better quality compared to joint or partner commissioned evaluations. See graph on next page.
- Evaluations with a DFAT member or M&E expertise on the team showed only marginally more adequate quality evaluations.

The results were not strong enough to corroborate the findings of the 2014 Review of program evaluations.

- The 2014 review found a stronger correlation between having a DFAT staff member and M&E expertise on the evaluation team and adequate or better quality evaluations.
- The 2014 review showed that for evaluation quality, the key is “everything in moderation”. Having either too few or too many total days, field days, evaluation questions and team members impacted adversely on evaluation quality.

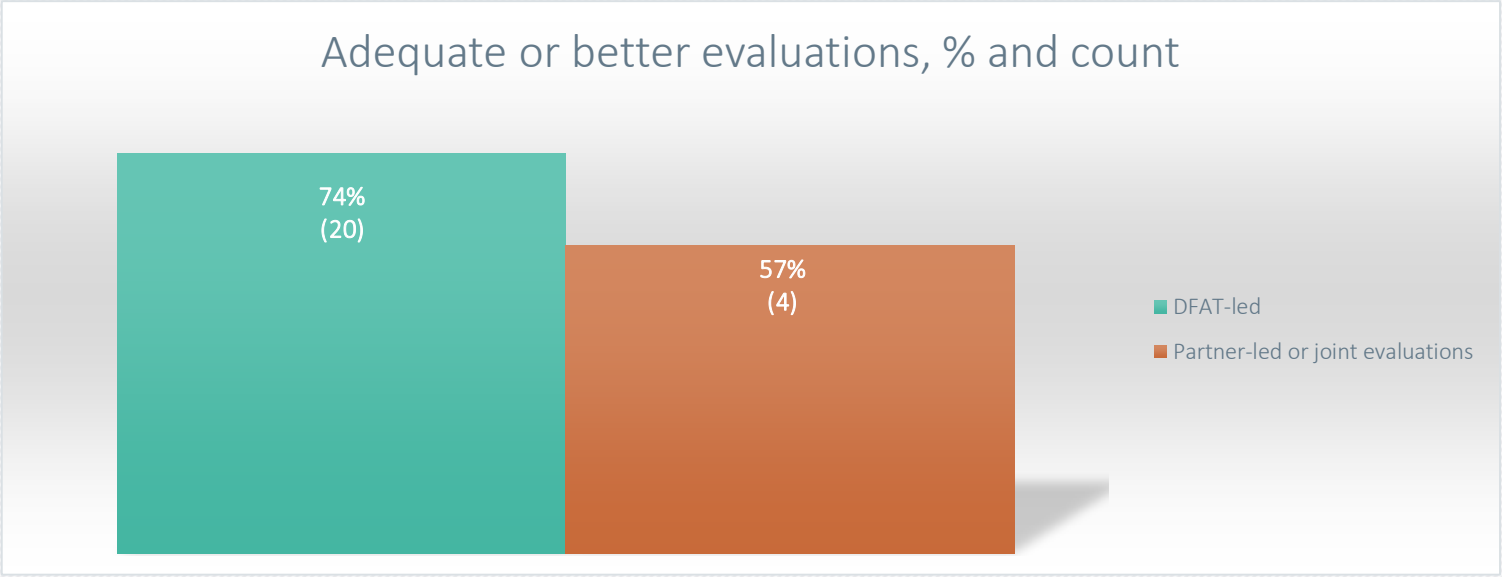
Qualitative analysis of reviewers’ comments to justify ratings for all nine quality criteria identified common themes related to evaluation practice that influenced the quality of 2017 program evaluations. See Tables on pages 43 - 46.

- These findings suggest that **good quality evaluations are more influenced by <u>how</u> they are planned and conducted** rather than any single underlying factor assessed above. Good quality evaluations have a clear management purpose; a scope and methodology that is appropriate to the resources available and the complexity of the investment; and strong oversight, including sound quality assurance processes, to ensure evaluation reports meet the DFAT M&E Standards. This led to the conclusion that evaluations that are fit for purpose and are actively managed by DFAT are more likely to be good quality.

FINDING 3: Evaluations that are fit for purpose and actively managed by DFAT are more likely to be good quality

DFAT COMMISSIONED VERSUS OTHER EVALUATIONS

- In 2017, evaluations commissioned by DFAT were more likely to be adequate or better quality compared to joint or partner-commissioned evaluations. A similar trend was seen in 2014 although the difference then was more pronounced. In 2014 87% of DFAT-commissioned evaluations were rated as adequate or better quality compared to 58% of joint or partner-commissioned evaluations.
- It is possible that for joint or partner-commissioned evaluations, DFAT devolves responsibility for evaluation quality to our partners and does not engage as actively compared to DFAT-commissioned evaluations.



FINDING 3: Evaluations that are fit for purpose and actively managed by DFAT are more likely to be good quality

The Review undertook qualitative analysis of reviewers' comments to justify quality ratings and identified the characteristics of stronger and weaker evaluation reports. These are highlighted in the table below.

CHARACTERISTICS OF STRONGER AND WEAKER EVALUATION REPORTS

STRONGER EVALUATION REPORTS	WEAKER EVALUATION REPORTS
• The report is easy to read, well structured, succinct and coherent. • The report has a clear line of sight between evaluation questions, findings, sources of evidence, conclusions and recommendations. • Findings are substantiated by a range of sources of evidence and reflect systematic and appropriate analysis and interpretation of the data, including identification of gaps and limitations. • The report uses appropriate methods and language to convince the the reader of the findings and conclusions e.g evidence tables, text boxes. • Complex issues are well explored - including enabling and inhibiting factors contributing to the program's progress/success or emergent challenges and opportunities - and appropriate solutions are proposed. • Conclusions and recommendations are logical and strategic and clearly take into account the views of a variety of stakeholders. • Annexes are used appropriately to support findings and analysis with additional information.	• The evaluation does not fully deliver against TORs and/or evaluation plan and does not provide an explanation of why. • Findings are presented as "facts" and are not well supported by adequate evidence from a range of sources, or are supported by contradictory claims. • Findings are largely descriptive and do not provide the reader with insight into why aspects of the program did or did not work well. • Findings largely draw on output data or there is little analysis of attribution or contribution to higher order outcomes. • The role of context or emergent risks to program performance is not analysed. • Failure to mention, draw on, or assess the quality of M&E data to substantiate findings. • The independence of the report is questionable, e.g. team leader is involved in the management or oversight of the program. • Recommendations do not flow logically from findings and conclusions.

FINDING 3: Evaluations that are fit for purpose and actively managed by DFAT are more likely to be good quality

We also undertook qualitative analysis of reviewers' comments to justify quality ratings for each major quality criterion to identify common themes. The tables below show examples of good practice and common weaknesses by major quality criteria.

GOOD PRACTICE AND COMMON WEAKNESSES BY QUALITY CRITERIA

PURPOSE AND SCOPE

- ✓ The purpose, objectives and primary users of the evaluation is clear.
- ✓ Matches evaluation resources, time, methods and skills of the team with the purpose and questions of the evaluation.
- ✓ Clearly outlines the roles and responsibilities of each team member and DFAT management.
- ✓ Includes consideration of the needs of women, and people living with disabilities and other disadvantaged groups in data collection methods.
- ✓ Matches the evaluation scope, methods and resources with the complexity of the investment and conte

- ✗ Fails to deliver all elements of the evaluation and/or compromises evaluation quality due to insufficient time and resources allocated to the evaluation.
- ✗ Does not explicitly link methods to individual evaluation questions.
- ✗ Fails to match evaluation resources to the evaluation purpose and objectives.

METHODOLOGY

- ✓ Outlines a methodology which is appropriate for the purpose and scope of the evaluation and proportionate to the value of the program.
- ✓ Clearly describes and justifies techniques for data collection and analysis and links these to evaluation questions.
- ✓ Identifies a range of methods and data sources to ensure triangulation of findings.
- ✓ Discusses limitations of the methodology and identifies mitigation strategies to address these.

- ✗ Does not describe sampling methods or its limitations for stakeholder interviews and program site visits.
- ✗ Does not identify appropriate methods to answer some of the questions.
- ✗ Fails to identify and discuss how ethical issues such as privacy, anonymity and cultural appropriateness were addressed.
- ✗ Does not include sufficient description of the methodology in the evaluation report or annexes.

FINDING 3: Evaluations that are fit for purpose and actively managed by DFAT are more likely to be good quality

GOOD PRACTICE AND COMMON WEAKNESSES BY QUALITY CRITERIA

ADEQUACY AND USE OF M&E

✓ Provides good detail about the investment's M&E systems and appropriately uses program monitoring data and reporting to substantiate findings.

✓ Assesses the adequacy of data used from the program M&E framework, including its strengths and weaknesses.

✓ Presents evidence which demonstrates the degree to which good performance information is available.

✓ Provides recommendations for strengthening shortcomings of the M&E system.

✗ Fails to present any data from investment M&E systems or explain why M&E information hasn't been used.

✗ Presents broad findings from M&E systems such as "the Results Management Framework shows that outcomes are on track for being achieved" without providing specific evidence to substantiate these findings.

✗ Fails to reference baseline data and does not assess progress against targets.

✗ Focuses on activities and outputs and does not provide evidence of progress or achievements against program outcomes.

CONTEXT

✓ Presents sufficient and relevant information to allow the reader to understand the relationship between the investment and its context (e.g. geographic, cultural, gender, social, political, economic and institutional)

✓ Provides good analysis of contributing, enabling and constraining factors impacting on program performance.

✓ Provides evidence of how the program has adapted to respond to changing circumstances and emerging opportunities.

✓ Analyses the policy environment and institutional factors, both in partner government and DFAT contexts that are facilitating or hindering progress.

✗ Does not analyse the role of the context and emergent risks to investment program performance.

✗ Does not analyse the impact the investment may have had on the context, e.g. institutional strengthening.

FINDING 3: Evaluations that are fit for purpose and actively managed by DFAT are more likely to be good quality

GOOD PRACTICE AND COMMON WEAKNESSES BY QUALITY CRITERIA

CREDIBILITY OF EVIDENCE

- ✓ Demonstrates a clear line of sight between evaluation questions, evidence, sources of data, analysis, findings and recommendations.
- ✓ Substantiates key findings with credible and convincing evidence and analysis and clearly identifies sources of data.
- ✓ Discusses gaps and limitations in the data and the impact on findings.
- ✓ Presents findings relevant to specific sub-groups (e.g.women, people with disability).

- ✗ Does not clearly and coherently present evidence from the range of data sources and methods used e.g. document reviews, site visits, interviews with key informants and beneficiaries.
- ✗ Does not discuss the implications of the findings sufficiently.
- ✗ Does not present the author's position clearly and unambiguously.
- ✗ Does not discuss alternative points of view.

RECOMMENDATIONS

- ✓ Recommendations flow logically from the key evaluation findings.
- ✓ Outlines clear, relevant, targeted, feasible actions, which have been discussed with relevant stakeholders.
- ✓ Addresses an appropriate balance of strategic and operational issues.
- ✓ If implemented, the recommendations are likely to bring about the required changes.

- ✗ Misses opportunities for influencing strategic change, e.g. DFAT role in policy dialogue.
- ✗ Identifies recommendations that are too broad so DFAT can agree to implement them without making specific changes needed to improve the program, e.g. establishing a Project Steering Committee without detailing priorities to be addressed by the Committee.
- ✗ Includes too many recommendations, which affects their prioritisation and implementability.
- ✗ Fails to provide information on who (job titles/work group) is responsible and timeframes for responding to actions, and resource implications (human, financial or material costs).

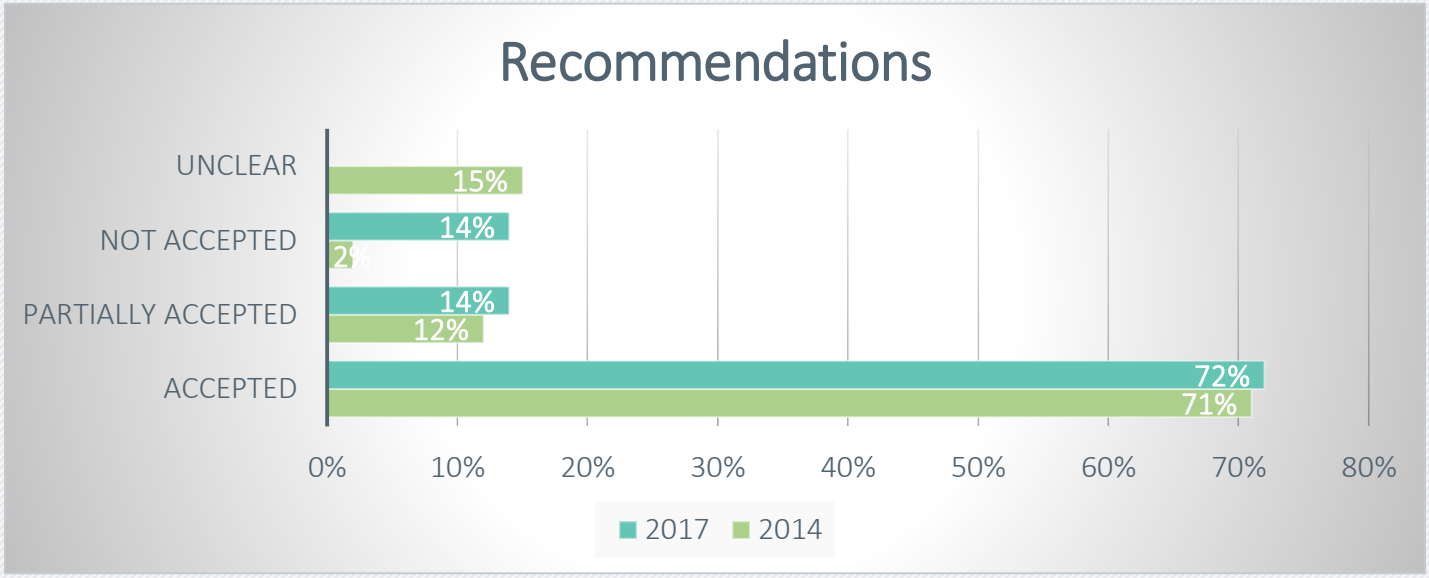
FINDING 4.The findings from program evaluations are being used to improve implementation and inform future aid designs

FINDING 4: The findings from program evaluations are being used to improve implementation and inform future aid designs

THE MAJORITY OF 2017 RECOMMENDATIONS WERE ACCEPTED

Analysis of the 32 available management responses showed similar positive results to 2014. 72% of recommendations were accepted, 14% partially accepted and 14% not accepted. This compares to 71% accepted in 2014, 12% partially accepted and 2% not accepted with 15% remaining unclear.

The majority of those recommendations not accepted were due to circumstances beyond the program's control e.g. a second phase not being implemented.



FINDING 4: The findings from program evaluations are being used to improve implementation and inform future aid designs

PUBLICATION AND MANAGEMENT RESPONSE RATES HAVE STRONGLY IMPROVED

The revised DFAT Aid Evaluation Policy requires that all evaluations identified in annual aid evaluation plans are published with a management response within three months of their completion.

The success of the revised policy can be seen in the high publication rate and management responses for evaluations identified on the 2017 Annual Evaluation Plan. This is a major increase on previous years, as illustrated in the graph below.

- At the end of 2017, 41 out of 43 evaluations on the revised evaluation plan were published and 39 of these included a management response. Prior to the new policy, only 38% were published and of these only around half had a management response (2016 ODE Review of Evaluations).
- These results indicate that the revised evaluation policy has laid the foundation for evaluations to be better used to their full potential.

Note: The measures before and after the introduction of the new evaluation policy are also different. Prior to introduction of the evaluation policy in 2016, the figures represent percentage of **completed** evaluations that were published and included management responses. The figures in 2017 represent the percentage of evaluations **identified** on the 2017 annual evaluation plan which were completed/published and of those the percentage that had a published management response. The 2017 figures also include ODE strategic evaluations which were otherwise not included in the Review of Program Evaluations.

FINDING 4: The findings from program evaluations are being used to improve implementation and inform future aid designs

USE OF 2017 EVALUATIONS SURVEY

In July 2018, a brief, qualitative survey was sent to all program areas responsible for managing the 2017 program evaluations to gauge how evaluations have been used.

Each program area was sent a list of evaluations that they had been responsible for managing and their accompanying recommendations, and two brief open-ended questions:

1. 'How have agreed recommendations been used to improve programs or inform future programming?'

2. 'If any agreed recommendations haven't yet been implemented, identify the factors that have hindered the uptake of the recommendations.'

The survey covered all 2017 program evaluations with recommendations and a management response. The survey received a 97% response rate.

The survey responses indicated that recommendations have largely been or are being implemented.

— Responses included many examples of how recommendations had been used to improve existing investments or inform the design of the next phase of existing investments or related new investments.

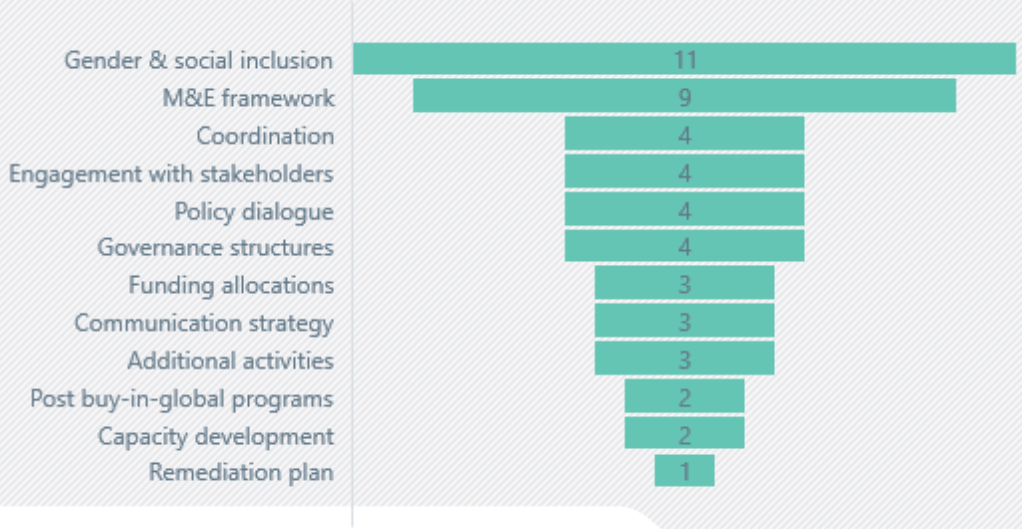
FINDING 4: The findings from program evaluations are being used to improve implementation and inform future aid designs

USE OF 2017 EVALUATIONS – TO IMPROVE EXISTING PROGRAMS

Common themes for recommendations being implemented were in the following areas:

- **improving gender and social inclusion,** e.g. Australia's Education Partnership with Indonesia – conducted a gender and disability analysis for the Indonesian education sector
- **strengthening M&E systems,** e.g. Australia Afghanistan Community Resilience Scheme – has now recruited an M&E advisor to develop a Performance Assessment Framework.

The graph below shows other areas in existing investments that are being strengthened through the implementation of 2017 program evaluation recommendations (Note: frequencies from qualitative data – number of evaluations that have already implemented recommendations related to these areas).



FINDING 4: The findings from program evaluations are being used to improve implementation and inform future aid designs

USE OF 2017 EVALUATIONS –

TO INFORM NEXT PHASES OR NEW INVESTMENTS

The recommendations from thirteen 2017 program evaluations are being/have been used to inform the next phase of existing investments or related new investments. Some examples include:

- Fiji Community Development Program (FCDP) – Recommendations used to inform the delivery of DFAT's assistance to civil society programs after FCDP ended, moving from a standalone civil society support program to an integrated approach under the Fiji Program Support Facility and Pacific Women – ensuring different types of funding and capacity support.

- Strengthening Pre-Service Teacher Training in Myanmar – Recommendations implemented through the phase II design including establishing a steering committee for the program.

- Integrated Coastal Management Program in Vietnam – The evaluation's recommendations have informed the design of the new GIZ funded 'Mekong Delta Climate Resilience Programme', with a focus on strengthening the regional coordination of all 13 Mekong Delta provinces.

FINDING 4: The findings from program evaluations are being used to improve implementation and inform future aid designs

USE OF 2017 EVALUATIONS – BARRIERS TO IMPLEMENTING RECOMMENDATIONS

Common reasons given for not implementing recommendations include:

- **resource constraints** including program budgets, staffing levels, time limitations
- **program discontinuation** including if a second program phase was not implemented. In some cases however, program discontinuation was the implementation of a recommendation (e.g. Palau Cleared Ground Demining Project).

Other barriers included: responsibility for implementing some of the recommendations rested with an implementing partner or partner government and action had not yet been taken; delays in the next phase of the investment being implemented; and "agreed in principle" recommendations reliant on policy changes, which have not yet taken place.

RECOMMENDATIONS

Strengthening awareness of DFAT evaluation requirements

The Review found that some 2017 program evaluations did not meet the requirements of the DFAT Aid Evaluation Policy and related guidance. An evaluation, as defined by DFAT, is an independent, systematic and in-depth assessment of an ongoing or completed investment/group of investments.

- Better awareness and understanding of evaluation requirements under the revised Aid Evaluation Policy is required to ensure effective leadership by senior managers and appropriate evaluation identification and management by operational staff.

To ensure that evaluation requirements are fully met, it is recommended that ODE:

1) engage more closely with Divisions on their consideration of evaluations to be nominated for the DFAT annual aid evaluation plan

2) identify ODE contact officers for each relevant Division to provide guidance on evaluation requirements and support evaluation capability.

Improving the quality of program evaluations

The Review found that the overall quality of program evaluations was satisfactory but there was room for improvement.

- Although 71% of evaluations were assessed as adequate or better quality, overall quality of evaluations had declined a small amount when compared to 2012 and 2014 Reviews.
- Quality related to the design elements of evaluations such as scope and methodology have improved over time but quality related to the "core essentials" or execution of evaluations such as use of program monitoring data and credibility of evidence and analysis have declined since 2012.
- Use and quality of investment monitoring systems was the weakest performing criteria and showed a 21% decline in quality compared to 2014.
- The Review also found that a higher proportion of inadequate evaluations were found amongst investments valued > $100 million.

To improve the quality of program evaluations, it is recommended that ODE:

3) review the terms of reference, evaluation plans and draft reports for evaluations of investments valued at, or greater than, $50 million.

Influencing better monitoring practice

The Review found there is a need to strengthen the use and adequacy of investment monitoring systems to ensure that evaluations are informed by robust and credible data.

— These findings suggest that DFAT needs to better monitor the establishment, quality and implementation of program M&E systems and play a more active role in influencing better monitoring practice amongst our implementing partners.

— Common themes and issues identified as part of this Review's qualitative analysis for use and adequacy of M&E systems (see page 37) could be used to guide where to best place efforts to improve investment monitoring systems.

To improve M&E systems, it is recommended that ODE:

4) liaise with the Contracting and Aid Management Division to consider options for strengthening investment monitoring systems to deliver more robust and credible data.

Australian Government

Department of
Foreign Affairs and Trade