Australian Government

Department of Foreign Affairs and Trade

Quality of Australian aid operational evaluations

Office of Development Effectiveness

June 2014

ISBN 978-0-9874848-2-6

Published by the Department of Foreign Affairs and Trade, Canberra, 2014.

This document is online at www.ode.dfat.gov.au

Disclaimer: The views contained in this report do not necessarily represent those of the Australian Government.

For further information, contact:

Office of Development Effectiveness
Department of Foreign Affairs and Trade
GPO Box 887
Canberra ACT 2601
Phone (02) 6178 4000
Facsimile (02) 6178 6076
Internet www.ode.dfat.gov.au

Office of Development Effectiveness

The Office of Development Effectiveness (ODE) at the Department of Foreign Affairs and Trade builds stronger evidence for more effective aid. ODE monitors the performance of the Australian aid program, evaluates its impact and contributes to international evidence and debate about aid and development effectiveness.

Visit ODE at www.ode.dfat.gov.au

Independent Evaluation Committee

The Independent Evaluation Committee (IEC) was established in mid-2012 to strengthen the independence and credibility of the work of the ODE. It provides independent expert evaluation advice to improve ODE's work in planning, commissioning, managing and delivering a high-quality evaluation program.

Foreword

Independent 'operational' evaluations are an essential part of the Australian aid performance management and reporting system. Good evaluations can inform the direction, design and management of the aid program. They also play an important accountability role, providing an independent perspective on the quality and results achieved through the Australian aid program. It is therefore important to periodically take stock of the quality, credibility and utility of these operational evaluations to make sure the Australian Government is getting it right.

The Office of Development Effectiveness (ODE) at the Department of Foreign Affairs and Trade is uniquely placed for such an undertaking. ODE's remit is to build stronger evidence for more effective aid in order to assist the continuous improvement of the Australian aid program. ODE draws its evidence mainly from its own evaluations of Australian aid and its analysis of aid performance systems.

This *Quality of Australian aid operational evaluations* review is an important piece of work. It casts a broader net to examine the bulk of the aid program's independent evaluations—i.e. those commissioned directly by aid initiative managers. It confirms that the majority of operational evaluations are credible. While further improvements in the quality of operational evaluations are possible, they do provide robust evidence for the performance of the Australian aid program.

This is the first quality review of operational evaluations overseen by the Independent Evaluation Committee. I hope such reviews will be undertaken on a regular basis.

I commend the report to you.

Jim Adams

Chair, Independent Evaluation Committee

Abbreviations

ACD	Contracting and Aid Management Division of the Department of Foreign Affairs and Trade
AusAID	the former Australian Agency for International Development[1]
CSO	civil society organisation
DFAT	Australian Government Department of Foreign Affairs and Trade
IEC	Independent Evaluation Committee
M&E	monitoring and evaluation
NGO	non-government organisation
ODE	Office of Development Effectiveness
OECD–DAC	Organisation for Economic Cooperation and Development—Development Assistance Committee
PEPD	the former Program Effectiveness and Performance Division[2] of AusAID/DFAT
PNG	Papua New Guinea
QAI	quality at implementation
UN	United Nations

1 At the time the reviewed evaluations were completed, and during the first part of this review, most of the Australian aid program was administered by AusAID, an executive agency. AusAID was integrated into the Department of Foreign Affairs and Trade (DFAT) in November 2013.

2 Renamed Contracting and Aid Management Division (ACD) in February 2014.

Contents

Acknowledgments

The Office of Development Effectiveness (ODE) would like to thank all those who contributed to this review.

The review team consisted of Nick Chapman (team leader), Hugh Goyder and Rob Lloyd from ITAD Ltd. The core DFAT management team for the review was led by Sam Vallance and Jo Hall from ODE, with Penny Davis and Simon Ernst from the former Program Effectiveness and Performance Division. The ITAD review team collected and analysed the data from the evaluations, while a collaborative approach was taken to the design of the review, the interpretation of the findings and framing of recommendations, and the drafting of this report. The Independent Evaluation Committee provided technical oversight.

The review team would like to thank the independent evaluators and DFAT managers who made time to be interviewed for the review, and the peer reviewers who provided feedback on the draft report.

Executive summary

The Office of Development Effectiveness (ODE) at the Department of Foreign Affairs and Trade (DFAT) builds stronger evidence for more effective aid. ODE conducts in-depth evaluations of Australian aid and analysis of aid performance management and reporting systems, to assist the continuous improvement of the Australian aid program.

Evaluation of the Australian aid program is undertaken at several levels and managed by different areas within DFAT. ODE evaluations typically focus on strategic issues or cross-cutting themes, and often entail cross-country comparison and analysis. ODE (under the guidance of the Independent Evaluation Committee) publishes only five or six evaluations each year. The vast bulk of DFAT's independent evaluations are commissioned by the managers of discrete aid initiatives. These are termed 'operational' evaluations to distinguish them from ODE evaluations and performance audits undertaken by the Australian National Audit Office. DFAT policy requires that all significant aid initiatives undergo at least one independent evaluation during their life cycle.

The 87 independent operational evaluations managed by program areas and completed in 2012 are the subject of this review by ODE. This review assesses the quality of the evaluations, considers underlying factors influencing evaluation quality and utility, and provides recommendations for improving evaluation quality and utility. (It was beyond the scope of this review to look at the quality or performance of the aid initiatives themselves.)

The timing for this review is opportune, as several recent developments—the introduction of the *Public Governance, Performance, and Accountability Act 2013*, the integration of the former Australian Agency for International Development (AusAID) into DFAT, and the planned simplification of the aid management system—open up additional opportunities for DFAT to take a lead role in demonstrating high-quality evaluation practice across the Australian Government, and to consider the role of the aid evaluation function within DFAT.

A companion ODE report, *Learning from Australian aid operational evaluations*, synthesises the findings from the evaluations that were found in this review to contain credible evidence and analysis, to provide lessons for improving the effectiveness of the Australian aid program.

Findings on quality and credibility of evaluations

Coverage of the aid program is satisfactory

Compliance with mandatory operational evaluation requirements was satisfactory. The initiatives evaluated were diverse in terms of value, sector and geographic region and can be considered broadly representative of the overall Australian aid program.

About half of the completed operational evaluations were published. While this represents a significant increase from previous publication rates, there is still room for improvement. There is also

scope to improve the accessibility of published evaluations. These improvements would provide greater transparency and support the development of a culture of learning from Australian aid evaluations.

The majority of evaluations are credible

The overall credibility of the evidence and analysis in the evaluation reports was satisfactory in 74 per cent of cases. Most evaluations satisfactorily assessed the relevance and effectiveness of the aid initiative, which indicates that operational evaluations are generally a robust source of evidence about the effectiveness of the Australian aid program.

Report quality could, however, be improved by a stronger focus on analysis of the extent to which the aid initiatives contribute to the outcomes observed and on the influence of context on initiative performance.

Evaluation quality could also be improved by taking a broader view of efficiency, including more substantive engagement with broader issues of value for money such as cost-efficiency or cost-effectiveness. However, a full value for money analysis may be beyond the scope of most evaluations.

The quality of the evaluations' assessments of initiative sustainability and advancement of gender equality could also be improved.

Up to a point, higher initiative value corresponded to better evaluation quality. However, it was concerning that a higher than average proportion of evaluations for very large initiatives ($100 million or greater value) were found to be to be of inadequate quality. This may in part reflect the complexity of very large initiatives, and suggests a need to focus on improving quality for these important evaluations.

It is also worth noting that, overall, the quality of evaluations managed wholly by the former AusAID was found to be at least as good as that of joint evaluations led by partners, and that this was generally achieved with fewer resources.

The design of evaluations could be improved

Greater attention to the evaluation design phase (evaluation terms of reference and evaluation plans) may also help strengthen overall evaluation quality. We found that most evaluations had a clear purpose but just over half did not assess the underlying logic of the intervention or had a weak assessment and/or did not adequately justify the evaluation methodology used. This in part reflects the absence of any specific guidelines requiring either an assessment of the underlying logic or a justification of methodology. It may also relate to the capacity of non-specialist staff to commission high-quality evaluations. This finding suggests a need to focus greater support and quality assurance efforts at this early stage.

Factors influencing evaluation quality and utility

The quality of an initiative's monitoring and evaluation system affects evaluation quality

The quality of evidence available to an evaluation team will depend in part on the quality of the initiative's monitoring and evaluation (M&E) system, which is intended to generate performance information over the life of the initiative. We were unable to determine the precise nature and

strength of the relationship between the quality of M&E systems and evaluation quality due to insufficient data; however, a separate ODE study is proposed for 2014 on the quality of initiative M&E systems.

Good evaluations require investment of funding, time and human resources

We found a positive correlation between evaluation quality and evaluation duration. Our findings also suggest that evaluation quality is influenced by the level of resourcing provided; however, we were limited by patchy data on evaluation cost. We found that, on average, more resources tended to be applied to evaluating larger initiatives and that, up to a point, higher initiative value corresponded with better evaluation quality.

Evaluation teams of three or four members (not including any Australian aid program staff involved) were more likely to produce adequate quality evaluations than teams of one or two, or teams of five or more. This and other evidence suggests that the expertise covered in the evaluation team is a key factor contributing to evaluation quality. This would usually include strong technical expertise in evaluation, good interpersonal skills, sector knowledge and, to a lesser extent, understanding of the country or regional context.

The evidence suggests that it is worthwhile for aid initiative managers to invest time and effort in building a strong relationship with the evaluation team. We also found that involvement of Australian aid program staff in delivering an evaluation can have numerous benefits, but that this involvement needs to be carefully defined so as not to compromise the independence of the evaluation.

Capacity to effectively manage evaluations is stretched

A central feature of the department's evaluation policy is mandatory evaluation of all significant aid initiatives. As a consequence, a large cohort of program staff are required to commission and manage evaluations as part of their normal program management duties. In recent years there have been significant efforts across the department to boost the capacity of non-specialist staff to help deliver high-quality evaluations. Nevertheless, realistic expectations need to be maintained as to the degree of evaluation expertise and knowledge these staff can or should acquire. Several interviewees suggested that, given the volume of evaluations undertaken, the evaluation capacity within the department is particularly stretched. Some also suggested that evaluation numbers are beyond the optimum for performance management and learning, and do little to assist the development of a department-wide culture of learning from evaluations.

These issues have already been recognised and are being addressed in several ways:

- In early 2012 the department's evaluation policy was revised, reducing the number of mandatory operational evaluations by approximately half to only one during the lifetime of each aid initiative. Further changes planned for mid-2014 will raise the financial threshold for aid initiatives requiring mandatory evaluation and will reduce evaluation numbers by a further 42 per cent.
- Country and regional program evaluation plans are being introduced to help improve the allocation of resources and skilled staff.
- The responsibility for monitoring and reporting on operational evaluations has shifted to ODE, marking a move towards consolidation of aid evaluation expertise within the department.

Evaluations with a clear and immediate program management purpose are more likely to be adequately resourced and be of higher quality

Evaluations are commissioned for various purposes, including to drive improvement, to inform future programming decisions and to provide accountability. This is reflected in guidelines for operational evaluations, which allow a high degree of flexibility in evaluation timing and resourcing.

In terms of timing, evaluations can be undertaken during initiative implementation (an independent progress report) or at the close of an initiative (an independent completion report). We found that:

- more independent progress reports are undertaken than independent completion reports
- independent progress reports had a higher average level of resourcing than independent completion reports
- a higher proportion of independent progress reports had a clear purpose and were of satisfactory overall quality.

This may indicate that independent progress reports are more useful to aid managers in terms of providing evidence and analysis to inform critical programming decisions such as whether to extend an investment.

However, departmental evaluation guidelines, while not mandatory, do not encourage flexibility in scoping. They set out expectations that all operational evaluations will assess initiative performance against a set of standard quality criteria (relevance, effectiveness, efficiency, sustainability, impact, and gender equality). Our review found evidence that this can potentially lead to an evaluation scope that is too ambitious to be realistic or appropriate, and that this can negatively affect evaluation quality and utility. While the evaluations' assessments against the key aid quality criteria of relevance and effectiveness were generally strong, about half of the evaluations' assessments against the other criteria were weak and were sometimes completed in a perfunctory manner (especially gender equality). This suggests a need for clearer departmental guidance that operational evaluations should be scoped to meet the specific information needs of program areas.

Assessing impact is difficult

Half of the evaluations did not attempt to assess the long-term impact of the aid initiative. Where an assessment of impact was made, the quality of the majority of those assessments was weak. This raises questions about the appropriateness of including the assessment of impact as standard in operational evaluations, given that the impact of an aid initiative is difficult to assess until well after its completion and typically relies on a robust monitoring and evaluation system being in place across the lifetime of the initiative.

Rigorous assessments of end-of-program outcomes and of impact remain a high priority to inform learning and account for the results of public spending on aid. Special arrangements for commissioning and resourcing evaluations specifically designed to look at the long-term impact of aid initiatives should be considered, particularly for high-value investments and/or those that offer broader learning opportunities.

Lessons for DFAT evaluation commissioning areas to maximise evaluation quality and utility

The evidence from our review highlights the following specific lessons for DFAT evaluation commissioning areas to maximise evaluation quality and utility.

Lessons for DFAT evaluation commissioning areas

Start evaluation planning well in advance

1. Consider the timing of an evaluation when you are developing initiative monitoring and evaluation arrangements. Plan the timing of the evaluation so that it will be most useful for program management purposes. Use program-level evaluation planning to help with the allocation of resources and skilled staff.
2. Start planning the evaluation six months ahead of planned commencement. Adequate time is needed to develop good-quality terms of reference (seeking support from performance and quality managers if needed), contract the most suitable consultants, engage with the consultants in their evaluation planning and schedule fieldwork to allow access to key stakeholders. Options may be limited if there is insufficient lead time.

Focus on developing strong terms of reference as the basis for a good-quality evaluation

3. Using the aid quality criteria as a starting point, develop a limited number of key evaluation questions that address the most critical issues and management decisions related to the initiative. Prioritise these questions to ensure a focus on the things that really matter. Consider including assessment of the intervention logic or theory of change. (If the intervention logic is not clearly articulated in the initiative design or implementation documents then one of the first evaluation tasks should be to reconstruct the intervention logic.)
4. Allocate sufficient time for the evaluation. This should match the scope of the evaluation but for a good quality evaluation would typically be two to three months from when the consultants commence to when the evaluation report is finalised. In particular, allow enough days for the consultants to develop a strong evaluation plan with a methodology that is appropriate to the evaluation questions, and for fieldwork.
5. Consider the skills required within the evaluation team, and the number of evaluation team members needed to cover this range of skills. Evaluation teams should consist of people with technical evaluation expertise and strong interpersonal skills, in addition to sectoral expertise and, to a lesser extent, knowledge of the country or regional context.
6. Be clear about the roles of any DFAT staff involved in the evaluation.

Continue to actively engage with the evaluation team during the evaluation

7. Invest time and effort in building strong relationships with the evaluation team.
8. Debate contentious issues, but respect the independence of the evaluation. Allow the team leader to exercise judgment on participation of staff in meetings or interviews.

These lessons should be read in conjunction with departmental evaluation guidelines, particularly the DFAT Aid Monitoring and Evaluation Standards, which articulate expectations of the quality expected from a range of M&E products. The relevant standards relating to independent evaluations are included at Annex 5. The standards had not been formally adopted at the time the evaluations examined in this review were undertaken, but they were integrated into evaluation guidance in 2012. The Standards provide a useful resource for evaluation commissioning areas.

Our review also identified several examples of good-practice evaluation documents, which are discussed in Annex 6.

Recommendations and management response

Acknowledging the need to improve the evidence base for effective aid programming and the principles of simplicity, proportionality and value for money, this review makes the following recommendations for improving the quality and utility of operational evaluations of Australian aid.

Recommendations and specific management responses

Recommendation 1 DFAT should review arrangements (including responsibility and resourcing) for the following evaluation functions: › evaluation planning at program level, including prioritisation and resourcing of evaluations › support by dedicated evaluation staff for non-specialist evaluation managers, particularly for developing evaluation terms of reference and/or evaluation plans and for evaluation of high-value investments.	**Management response** Agreed. The Government is committed to improving the effectiveness and efficiency of Australian aid, and has introduced a new performance framework which will see funding at all levels of the aid program linked to progress against rigorous targets and performance benchmarks. Independent evaluations are an important component of this strong focus on results and value for money. The review findings have been used to inform the development of a new aid management architecture. This includes: › explicit evaluation plans in country/regional Aid Investment Plans and annual performance reports › the creation of a unit within ODE which, alongside performance and quality managers, provides evaluation support across the department. An evaluation tracking database has been established to assist the targeting of support to evaluations of high value, high risk or otherwise strategically important investments.
Recommendation 2 DFAT should make it explicit that the purpose of the evaluation guides the approach to that evaluation. Specifically: › operational evaluations should remain flexible in timing, with their scope and methodology purposefully designed to meet the specific information needs of program areas › consideration should be given to commissioning impact evaluations, especially of high-value investments and/or those that offer broader learning opportunities.	**Management response** Agreed. This review has been particularly helpful in shaping the evolution of the department's evaluation guidance. Specifically: › revised evaluation guidance will emphasise that evaluation purpose be the key guide to determining the overall evaluation approach › the department has raised the minimum value threshold for mandatory evaluations to encourage fewer, better quality evaluations › ODE will work with programs to identify areas where impact evaluation may be of strategic value. It is anticipated that this will result in ongoing collaboration between ODE and programs on a limited number of impact evaluations.
Recommendation 3 DFAT should monitor implementation of the policy requirement to publish all independent operational evaluations and should improve their public accessibility.	**Management response** Agreed. ODE will monitor and support the publication of completed evaluations. An online register with links to completed evaluations will be incorporated into ODE's webpage.

1 About this review

1.1 Operational evaluations

Independent evaluations[3] complement annual self-assessment processes such as 'quality at implementation' reporting, where program areas assess the progress of their initiatives. Good evaluations can inform the direction, design and management of the aid program. Independent evaluations also play an important accountability role in the aid program's performance management systems, providing an independent perspective on the quality and results achieved through the Australian aid program.

Most of the Australian aid program is managed by the Department of Foreign Affairs and Trade (DFAT), which in November 2013 took responsibility for overseas development assistance, previously administered by the Australian Agency for International Development (AusAID). Evaluation is undertaken at several levels and managed by different areas within the department.

Program areas in DFAT commission and manage independent evaluations of individual aid initiatives. It is these initiative-level 'operational' evaluations managed by program areas that are the subject of this quality review. Currently an independent evaluation is required for every monitored initiative (i.e. those that are valued over $3 million or have strategic or political importance) at least once over its life.

Other types of independent aid evaluations include:

- evaluations of broad strategic relevance undertaken by the Office of Development Effectiveness (ODE) in line with its evaluation policy and three-year rolling work program, under the oversight of the Independent Evaluation Committee established in 2012
- sector-wide evaluations occasionally undertaken by thematic areas.

The Pacific program is also partnering with the Overseas Development Institute (ODI) to explore approaches to impact evaluations for specific aid initiatives.

1.2 Objectives

The objective of this review is to promote good-quality independent evaluations of aid initiatives. The review assesses the quality of the 87 independent operational evaluations completed in 2012[4] to identify actions that should be taken to improve the quality and utility of independent evaluations of Australian aid.

---

3 DFAT defines independent evaluations as evaluations that are led by a person or team external to the program area where there is no undue influence exercised over the evaluation process or findings.

4 The date on the final evaluation report falls between 1 January and 31 December 2012.

The review seeks to answer the following questions:

1. What are the basic characteristics of different levels of independent evaluation in the aid program and the history and nature of independent evaluation at the initiative level?
2. To what degree do independent operational evaluations provide a credible source of evidence for the effectiveness of the Australian aid program?
3. What are the major strengths and weaknesses of independent evaluations of Australian aid initiatives?
4. What are the factors that contribute to their quality and utility?
5. What actions should be taken to improve the quality and utility of independent operational evaluations?

The structure of this report follows these key evaluation questions.

A second ODE report, *Learning from Australian aid operational evaluations*, synthesises the findings from the 64 evaluations that were found in this review to contain credible evidence and analysis, to provide lessons for program design and management and for the broader community of development and aid actors.

The terms of reference covering both reviews are at Annex 1.

1.3 Approach

The review was undertaken from May 2013 to February 2014 by a team of consultants from ITAD Ltd and managed by ODE in partnership with DFAT's former Program Effectiveness and Performance Division.[5] The review was undertaken at the request of the Independent Evaluation Committee, which also provided technical oversight.

The review was based primarily on a desk assessment of evaluation reports and interviews of a sample of independent evaluators and DFAT aid program staff. An overview of the methodological approach taken for this review is provided here, with full details at Annex 2.

The review team assessed each of the 87 independent operational evaluations completed in 2012 against 15 quality criteria:

- Nine of the criteria are general evaluation quality standards outlining features that a good evaluation ought to have (for example, clear purpose and scope, appropriate methodology and credible evidence and analysis).
- Six of the criteria are standard Australian aid quality criteria. Under current guidelines, each operational evaluation should assess initiative relevance, effectiveness, efficiency, impact, sustainability and advancement of gender equality. We looked at how well these six criteria were applied.

[5] The initiative-level independent evaluation support function sat with the Program Effectiveness and Performance Division until February 2014, when it moved to ODE. At the same time, the Program Effectiveness and Performance Division was renamed Contracting and Aid Management Division. By that time the preparation of this report was in its final stages.

The team analysed the extent to which there were correlations between evaluation quality and the characteristics of the initiatives evaluated (for example, value or sector), or between evaluation quality and evaluation characteristics (for example, type or length of evaluation).

The team interviewed a total of 27 people—independent evaluators, Australian Government aid initiative managers and senior aid program executives—and analysed the interview responses. Taking an 'appreciative inquiry' approach (emphasising understanding and learning from positive experiences) the interviews sought to identify the factors contributing to stronger or weaker evaluations.

The team identified examples of good practice evaluation products (terms of reference, evaluation plans, and evaluation reports) to provide learning material for use by aid initiative managers and evaluators.

The team then triangulated between the various forms of evidence and analysis to identify the key conclusions and recommendations presented in this report.

2 Independent evaluation of Australian aid

This chapter outlines policy requirements and support for operational evaluations within the Department of Foreign Affairs and Trade (DFAT), including recent and planned changes. It also provides background on the introduction of the *Public Governance, Performance and Accountability Act 2013* and the evaluation-related recommendations of the *2011 Independent Review of Aid Effectiveness*. The timing of this review is opportune, as these developments open up additional opportunities for DFAT to take a lead role in demonstrating high-quality evaluation practice across the Australian Government and to consider the role of the evaluation function within DFAT.

2.1 Evaluation policy and support

Policy requirements for operational evaluations in 2012

The operational evaluations reviewed in this report were completed in 2012. In early 2012, the evaluation policy of the former AusAID was as follows:

- All monitored aid initiatives (those valued over $3 million or with strategic or political importance)[6] required evaluation at least once every four years. In practice this meant that many initiatives required evaluation during implementation and at completion.
- All evaluations were expected to assess the initiative against eight evaluation criteria: relevance; effectiveness; efficiency; impact (where feasible); sustainability; monitoring and evaluation; gender equality; and analysis and learning. The first five of these criteria are based on the internationally agreed aid effectiveness criteria identified by the Organisation for Economic Cooperation and Development—Development Assistance Committee (OECD–DAC), while the final three criteria are specific to the Australian aid program.
- All evaluations were required to rate the initiative against the above criteria (except for impact). A six-point scale was used, with 1 indicating that an initiative was very low quality in relation to a particular criterion and 6 indicating that it was very high quality.[7]
- All evaluations were subject to formal peer review.
- A formal management response from the evaluation commissioning area was required.

An updated policy came into effect in March 2012 and detailed guidance on implementing the policy was released in December 2012. The new evaluation policy stated that:

---

6 In 2012–13 approximately 68 per cent of all aid funding administered by the former AusAID was monitored. The types of initiatives that are not monitored are administrative support activities, Australian Civilian Corp deployments, and humanitarian assistance and emergency response where the duration is less than 12 months. Multilateral core funding is covered by other quality processes, and direct appropriations to other government departments are not covered.

7 The full rating scale is 1 = very poor quality, 2 = poor quality, 3 = less than adequate quality, 4 = adequate quality, 5 = very good quality, and 6 = very high quality.

- All monitored aid initiatives required evaluation once over their life at the best time for program purposes.
- All evaluations were expected to consider the following six standard Australian aid quality criteria: relevance; effectiveness; efficiency; sustainability; impact; and gender equality. If a particular criterion had already been addressed through another assessment process (such as a partner-led evaluation), it did not need to be included in the evaluation.
- All evaluations were expected to rate the initiative against the chosen criteria using the existing six-point scale.
- Peer review was recommended but no longer mandated.
- A formal management response by the evaluation commissioning area was required and had to be published.

Under this updated policy it was expected that approximately 110 operational evaluations would be conducted in 2012.[8]

The new evaluation guidance also introduced the AusAID Monitoring and Evaluation Standards. These standards—now the DFAT Aid Monitoring and Evaluation Standards—provide guidance for staff and evaluators on what good-quality evaluation products look like. The relevant standards (Standard 4: Terms of Reference for Independent Evaluations, Standard 5: Independent Evaluation Plans, and Standard 6: Independent Evaluation Reports) are included at Annex 5 of this report. In addition, the updated evaluation guidance provided some new information on conducting impact evaluations. It encouraged initiative managers who were considering an impact evaluation to refer to a discussion paper on this topic released by the Office of Development Effectiveness (ODE) in September 2012.

At each iteration of the evaluation policy, joint or partner-led evaluations have been encouraged for initiatives that are co-financed or implemented through partners. These evaluations aim to share learning across partners and to avoid overburdening partner governments and beneficiaries with multiple evaluation processes.

Departmental integration and simplification of the aid management system

In November 2013 AusAID ceased as an agency and most of Australia's aid program is now delivered by DFAT. An integration process is under way towards a transformed DFAT that aligns and implements foreign, trade and aid policies and programs in a coherent manner. In the transition process the department recognises that high-quality evaluations remain central to an effective aid program that advances Australia's national interests.

A business model review conducted by AusAID immediately before integration found a need to simplify and streamline the aid management system in order to improve both effectiveness and efficiency. Departmental integration has made the planned simplification of the aid management system more urgent.

As a result, further revisions of the evaluation policy are being undertaken. The new evaluation policy will not come into effect until July 2014 and many details are still being decided, but the following parameters have so far been determined:

---

[8] In the 2011–12 financial year there were 588 monitored initiatives, with an average duration of 5.3 years. At least one evaluation was required during the life of every initiative so, assuming an even split of evaluations per year, it was expected that 111 initiatives would be evaluated in 2012.

- All aid initiatives over $10 million will require evaluation once over their life at the best time for program purposes.
- All country programs are required to develop evaluation pipeline plans and to update them annually.

Under this revised policy it is anticipated that approximately 64 operational evaluations will be completed each year, representing a significant reduction since early 2012 in the number of evaluations required.

Support for operational evaluations

There is currently no consistent or centralised approach to quality assuring operational evaluations. Program areas internally quality assure their own evaluation terms of reference, plans and reports. Many evaluation reports are peer reviewed, although the approach varies considerably. Under a previous centralised independent technical review process for operational evaluations, consultants reviewed more than 70 draft evaluation reports between 2008 and 2010, however this process was discontinued in 2011.

Program areas have designated performance and quality staff who coordinate evaluation planning and assist initiative managers to commission operational evaluations (as well as undertaking other quality processes). An active performance and quality network provides regular opportunities for these staff to share lessons and develop their knowledge and capacity in practical application of performance assessment and evaluation.

Several country programs also fund evaluation capacity building programs for their staff. Different program models have been used; however, the core aim of these internal programs is to build the capacity of DFAT staff to commission, assess and use high-quality monitoring and evaluation products for their initiatives. Some evaluation capacity-building programs also work with external contractors to build their ability to produce the high-quality monitoring and evaluation products that DFAT demands. In early 2014, eight country or regional programs had, were planning, or had completed an evaluation capacity building program.[9]

The department also provides support to program areas in all aspects of performance management. This includes formal training in monitoring and evaluation. The function of providing support and guidance to program areas commissioning operational evaluations, as well as monitoring and reporting on operational evaluations, was transferred from the Program Effectiveness and Performance Division to ODE in February 2014 to optimise coherence and efficiency.

2.2 Background

Public Governance, Performance, and Accountability Act

Since late 2010, the Australian Government has been reviewing the framework for financial accountability across the Public Service to improve performance, accountability and risk management. As a result the Public Governance, Performance and Accountability Act was introduced in 2013, with its key provisions to commence on 1 July 2014. This reform aims to transform the way

[9] The country/regional programs are the Philippines, Africa, Indonesia, Timor-Leste, Vanuatu, Pacific Regional, Samoa and Fiji.

that Australian Government agencies approach financial management, including recognition that performance of the public sector is more than financial. Requirements for performance management will be made more consistent across government agencies, with implications for how evaluation is conducted.

The aid program is ahead of other government programs in this regard given its long history of evaluation and established performance management systems and DFAT will be well positioned to take a lead role in demonstrating high-quality evaluation practice across the Australian Public Service. As the Act provisions come into place, it will be important for DFAT to maintain a high-quality and systematic program of evaluations. This review will provide a baseline for quality of evaluations and highlight underlying factors contributing to evaluation quality.

Independent Review of Aid Effectiveness

The 2011 *Independent Review of Aid Effectiveness* was a comprehensive review of Australia's aid program. To support that review, a study of independent evaluations of the aid program was commissioned, which found issues with evaluation quality and compliance (see Box 1).

Box 1 2011 Study of independent completion reports

Peter Bazeley's *Study of independent completion reports and other evaluation documents* rapidly reviewed evaluations from a four-year period from July 2006 to June 2010. It found that there was low compliance with agency evaluation requirements: 236 evaluations were identified by AusAID as having been completed in that period.

The study also found the following issues with the quality of evaluations:

- Independent completion reports were undertaken primarily for accountability purposes and not for learning or management.
- Approximately one-quarter of evaluation reports were poor quality and not publishable.
- More than 60 per cent of initiative monitoring and evaluation (M&E) systems were deemed 'less than satisfactory': they had inadequate or non-existent baselines, and provided insufficient or poor quality quantitative data on costs, intermediate impacts and outcomes.
- The average time allowed (23 days) was minimal given the evaluation expectations.
- Evaluations focused on low-level activity and process over strategic or higher-level outcomes/impact.
- There was poor discussion of efficiency, with a major focus on 'low-level activity management processes' rather than value for money.
- There was a narrow, variable interpretation of evaluation criteria.

Source: Peter Bazeley, Study of independent completion reports and other evaluation documents, 2011.

The *Independent Review of Aid Effectiveness* subsequently made a number of recommendations in relation to independent evaluation of Australian aid. It called for the establishment of the Independent Evaluation Committee (IEC) to oversee the work of ODE and allow for greater independence in evaluations. In 2012 the IEC was established. In its first year of operations it focused on the rolling program of evaluations conducted by ODE, quality assurance reports produced by ODE and ODE's first *Lessons from Australian aid* report (a synthesis of independent evaluations and quality assurance reports—also a recommendation of the *Independent Review of Aid Effectiveness*). The IEC requested this *Quality of Australian aid operational evaluations* review as its

first look at evaluations of the aid program beyond those of ODE. This review has been prepared under the oversight of the IEC.

The *Independent Review of Aid Effectiveness* also recommended an overhaul of the aid program's independent evaluation system (see Box 2). It called for an overhaul of the evaluation system with a view to 'undertake fewer, but higher quality independent evaluations and publish them all'. This has been progressed through revisions to the evaluation policy since 2012 (see section 2.1) and efforts to improve the publication rate.

Box 2 Evaluation conclusions of the 2011 Independent Review of Aid Effectiveness

The 2011 *Independent Review of Aid Effectiveness* found that:

> *While AusAID's self–rating system has taken great strides in recent years, the evaluation system, though equally important, is not working well, and requires an overhaul.*
>
> ***Undertake Fewer, but Higher Quality Independent Evaluations and Publish them all***
>
> *The compliance burden of the current system is too high. The number of projects that require an independent evaluation should be reduced.*
>
> *Every substantial project should have a completion report of some form or another. This should be a management responsibility, and it would be a management decision as to whether these reports were done independently or not.*
>
> *A smaller number of projects, themes and country programs should be chosen for mandatory independent evaluation.*
>
> *The aim should be to do about 10–20 of these independent evaluations every year. Maintaining an annual independent evaluation plan, implementing it and ensuring quality and publication would be the responsibility of ODE.*
>
> *Centralising quality control in this way is critical to improve report quality. This will help AusAID feel more comfortable about publishing them. The current system is decentralised, leading to variable quality.*
>
> *Under the new system… ODE would have to vouch for the quality of the evaluation, even if not endorsing the contents. ODE could also sign off on evaluation teams and ensure the quality of consultants eligible to be contracted to write evaluation reports.*
>
> Source: Independent review of aid effectiveness, pp. 297–298.

3 Characteristics of initiatives evaluated and evaluations

3.1 Evaluation coverage and characteristics of the initiatives evaluated

The level of coverage is satisfactory in terms of compliance with DFAT's evaluation policy, which requires an evaluation once in the life of every significant aid initiative. Of the 87 operational evaluations completed in 2012, six were 'cluster' evaluations covering more than one initiative, so the total number of initiatives actually covered by the evaluations was 94.[10] The expectation was that approximately 110 initiatives would be evaluated in 2012.

Of the 87 evaluations completed in 2012, 42 (48 per cent) were published externally on the Australian aid website as of 30 September 2013, which is a significant improvement on previous publication rates.

The 94 initiatives evaluated are diverse in terms of value, sector and geographic region, and can be considered broadly representative of the overall aid program. In summary:

- **Initiative value**: The total approved investment value represented by the 94 evaluated initiatives was approximately $3.2 billion.[11] Initiatives ranged in value from $200,000 to $212 million, with an average of $35 million. Sixty-eight per cent of initiatives were valued at less than $30 million, while there were nine initiatives valued at $100 million or more.
- **Sector**: The 94 evaluated initiatives cover all major sectors[12] (see Figure 1). In comparison to the total aid program for 2011–12,[13] education appears to be rather under-represented in the evaluations, and improved governance somewhat over-represented. While the broad strategic goal of effective governance (which includes the improved government, security and justice, and human rights sectors) represented 19 per cent of Australian Government aid investment in 2011–12,[14] the value of the evaluated initiatives included in this review that had effective governance as their primary strategic goal represented 43 per cent of the total value of the evaluated initiatives. This concentration is partly explained by the inclusion of evaluations of five very large initiatives

---

10 Some of the 94 initiatives may be only partially covered, as some evaluations covered only selected activities within an initiative.

11 Based on financial approval value over the life of the initiative under Financial Management and Accountability Regulations 9 and 10.

12 Only the small mineral resources and mining sector ($14 million expenditure in 2011–12) is not covered.

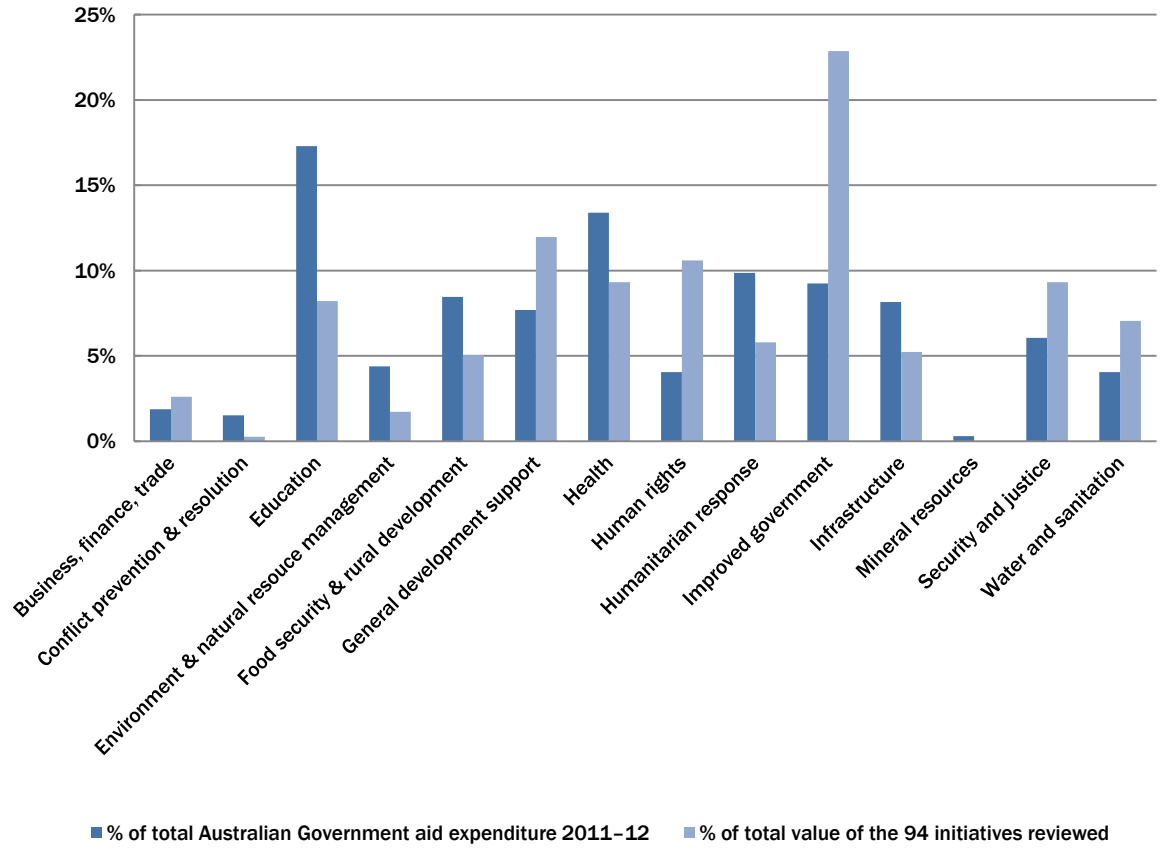
13 Australia's International Development Assistance Program 2013–14. Direct comparison between the value/expenditure of the Australian aid program and the 94 evaluated initiatives is difficult because the available figures for the former are for expenditure in a single year, whereas the available figures for the 94 evaluated initiatives are for total investment value over the life of the initiative (covering multiple years). The comparison presented here is indicative rather than definitive.

14 Australia's international development assistance—Statistical summary 2011–12, p. 5.

together valued at $755 million, or almost one-quarter of the total value of the 94 evaluated initiatives.

› **Country and region**: The 94 evaluated initiatives were spread across all major geographic regions covered by the aid program.[15] Papua New Guinea and Indonesia—Australia's two largest country programs by value—had the highest representation by total initiative value, with respectively 23 per cent and 15 per cent of the total value of the evaluated initiatives. There was also a significant proportion (16 per cent) of multicountry initiatives.

Figure 1 Indicative comparison by sector between overall aid program expenditure in 2011–12 and investment value[16] of the 94 evaluated initiatives



3.2 Characteristics of the evaluations reviewed

Eighty-seven independent evaluations of Australian aid initiatives were completed in 2012 and are included in our review. The characteristics of these evaluations are described below.

15 The pool did not include any evaluations for initiatives in Latin America, the Caribbean or the Middle East, although the evaluations of global or multicountry initiatives may have covered some of the Australian aid program's small amount of aid expenditure in these regions.

16 Based on financial approval value over the life of the initiative under Financial Management and Accountability Regulations 9 and 10.

Evaluation type

- Fifty-one of the 87 evaluations reviewed (59 per cent) were independent progress reports (including mid-term reviews), which were completed on average two to 2.5 years before the end of the initiative. Thirty-four (39 per cent) were independent completion reports, which were completed about one year before the end of the initiative. Two were ex-post evaluations completed around five years after the end of the initiative.[17]
- Six of the 87 evaluations reviewed were 'cluster' evaluations covering more than one initiative. Initiatives were evaluated as a cluster because they were contemporaneous and fell within the same sector and country or because they formed separate phases or types of support for the same activity.[18]
- Fifteen evaluations were joint evaluations led by partners rather than managed by the former AusAID alone. These evaluations were mostly led by the World Bank or the United Nations.

Evaluation cost

Evaluation budget details were available for only 39 of the AusAID-managed evaluations and do not include the cost of Australian Government aid program staff involvement where this occurred. For these 39 evaluations:

- The average cost was $90,000 but most evaluations cost less than this, with a median cost of $80,000. One-third cost more than $100,000.
- More resources tended to be applied to evaluating larger initiatives.[19] Evaluations of initiatives valued at more than $50 million cost on average $125,000, while evaluations for initiatives valued at $50 million or less cost on average $78,000.
- Independent progress reviews were on average more expensive than independent completion reports, at an average of $98,000 for the former compared to $77,000 for the latter.[20] This is most likely linked to the larger average team size for independent progress reviews (discussed below).

Reliable evaluation cost data was not available for the partner-led joint evaluations, but if overall length, number of fieldwork days and number of team members are used as proxy measures for resourcing, then the partner-led joint evaluations were, on average, better resourced than the evaluations managed by the former AusAID alone (discussed below).

Evaluation team

- Seventy-eight of the 87 evaluation reports had information on team size. The evaluation teams ranged in size from one to eight, with the majority having one, two or three members (17 per cent had a single evaluator and 60 per cent had two or three people).

---

17 The two ex-post evaluations were for the Agusan Del Sur Malaria Control and Prevention Project Community Trust Fund in the Philippines, and the Laos–Australia Basic Education Project (1999–2007). See Annex 6 for further information on these initiatives and their evaluations.

18 These 'cluster' evaluations are marked in the evaluation list in Annex 6.

19 A significant level of association was found, with a correlation coefficient of 0.53 between initiative value and evaluation cost.

20 Cost details were available for 26 independent progress reports and 12 independent completion reports.

- On average, independent progress reports had larger teams than independent completion reports. Twenty-four of the 45 independent progress reports with information on team size available (53 per cent) had teams of three or more, compared to only nine of the 31 independent completion reports (29 per cent) for which this information was available.
- Seventy-four different team leaders were used. One team leader completed three evaluations and five completed two evaluations.
- In just under half of the evaluations (42 of the 87), Australian Government aid program staff were also involved in conducting the evaluation. This typically included aid initiative or program managers, or in some cases specialist advisers (e.g. gender or sector specialists). However, as the extent of their role was not always stated, we did not include these staff members as team members in our analysis around team size.

Details were not readily available on the range of skills and experience or areas of specialisation of team members.

Evaluation duration

The length of the evaluations was measured in two ways: the total length from the start of the evaluation to the evaluation report submission; and the number of fieldwork days. Calendar days (the overall period of time for the work) rather than person or consultancy days were used, as this data was more consistently reported.

- For the 58 evaluations that had data available, the total length of most was three to six weeks, with a range from 10 to 270 days. Partner-led joint evaluations were longer on average (averaging 80 days compared to 37 for evaluations managed by the former AusAID alone). Some of these are major studies such as the evaluations of UN Delivering as One reform or the Afghanistan Reconstruction Trust Fund.
- For the 67 evaluations that had data available, fieldwork was on average 15 calendar days with a range from zero to 42 days. On average, partner-led joint evaluations had five more fieldwork days than evaluations managed by the former AusAID alone.

Assessments and ratings against standard Australian aid quality criteria

As discussed in section 2.1, across DFAT's aid performance management and reporting systems standard criteria are used for evaluating the quality of Australian aid. Of these, the following five criteria are based on the internationally agreed aid effectiveness criteria identified by the OECD-DAC: relevance, effectiveness, efficiency, impact[21] and sustainability. Two further criteria are also widely used within DFAT: monitoring and evaluation, and gender equality.[22] While there was some variation across the criteria, we found that most evaluations provided assessments against most of these criteria, with the notable exception of impact. The coverage and quality of these assessments is discussed in section 4.3.

Guidelines in effect in 2012 recommended the inclusion of numerical ratings for initiative performance against these criteria (except impact). A six-point rating scale is used, with 1 indicating very poor quality and 6 indicating very high quality. However, fewer than half of the 87 evaluations

[21] Used in some independent evaluations but not in 'quality at implementation' performance reporting.

[22] Under the evaluation guidance in place in early 2012, assessments were also sometimes made against an additional analysis and learning criterion. This review did not consider those assessments, as analysis and learning is not considered a core quality criterion and has since been dropped.

actually provided numerical ratings. The criteria were not covered equally: 38 evaluations provided ratings for effectiveness (44 per cent); 37 rated relevance, efficiency and sustainability (43 per cent); 35 rated monitoring and evaluation (M&E) (40 per cent); 33 rated gender equality (38 per cent); and only nine rated impact (10 per cent). Evaluation reports did not provide ratings against the standard Australian aid quality criteria for various reasons; most commonly it was because the terms of reference did not require ratings to be provided or because the evaluation was a partner-led exercise where no ratings were provided or a different rating system was used that was not comparable to the DFAT system.[23] Section 4.3 considers the robustness of the ratings.

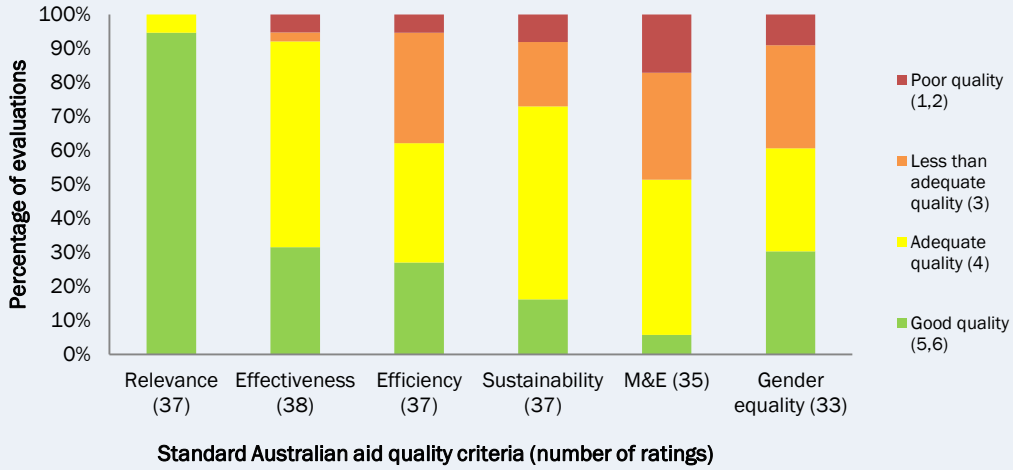
Box 3 What do the ratings tell us about the performance of aid initiatives?

While this review focuses on the quality (coverage and robustness) and, to a lesser extent, utility of the ratings provided in the evaluations, the question of what the ratings actually tell us about the performance of Australian aid initiatives is also of interest.

That fewer than half of the evaluations provided numerical ratings limits our ability—and the ability of areas across the department—to make findings based on evaluation ratings with a high degree of confidence. Given that caveat, the ratings data can still provide an overview of the assessed performance of these initiatives (see Figure 2). In summary:

- Relevance was rated as adequate or better in all rated initiatives, including all but two rated as either good or very high quality.
- Effectiveness was rated as adequate or better in 92 per cent of rated initiatives, but with the majority rated as adequate rather than good.
- Efficiency was rated adequate or better in 62 per cent of rated initiatives.
- Sustainability was rated as adequate or better in 73 per cent of rated initiatives.
- M&E was rated adequate or better in only 51 per cent of rated initiatives, with more initiatives rated as poor or very poor than any other criterion.
- Gender equality was rated adequate or better in 61 per cent of rated initiatives.
- Impact was rated in only nine evaluations (so is not included in Figure 2).

Figure 2 Standard Australian aid quality criteria—ratings data from the evaluations



[23] Evaluations that used a ratings system that was not comparable to the DFAT system were counted as 'not rated' for the purposes of this review.

4 Quality and credibility of evaluations

This chapter considers the degree to which operational evaluations provide a credible source of evidence for the effectiveness of the Australian aid program, and identifies the major strengths and weaknesses of operational evaluations.

Our review team assessed the quality and credibility of the 87 independent evaluations completed in 2012 against a total of 15 quality criteria (set out in Figure 3 below).

The 15 quality criteria include nine general evaluation quality standards. These are features that a good evaluation ought to have (for example, clear purpose and scope; appropriate methodology; and credible evidence and analysis). They draw on a selection from the DFAT Aid Monitoring and Evaluation Standards for good practice (see Annex 5), especially standard 6 relating to evaluation reports. These standards in turn draw on the internationally agreed OECD–DAC evaluation quality standards.

We have assessed each evaluation against these generic criteria without taking into account the intended purpose or utility of the individual evaluation. Evaluations can be undertaken for a range of purposes, as reflected in the variation in scope, design and resourcing. Some evaluations are designed to rapidly review the progress of an aid initiative within a narrowly defined scope, while others are intended to be comprehensive 'gold standard' evaluations. Aid managers have some freedom to scope evaluations according to their information needs. However, it was not within the scope of this review to assess whether the evaluations were fit for the purpose intended and we have therefore reviewed the quality of the evaluations against the nine generic evaluation quality standards. It is reasonable to expect all evaluations to at least minimally meet these standards, with the possible exception of the 'assessment of intervention logic' standard (as discussed below).

The review team also assessed the quality of each evaluation's assessment against six[24] standard Australian aid quality criteria (discussed in section 3.2). Under current guidelines, each operational evaluation should assess initiative relevance, effectiveness, efficiency, impact, sustainability and advancement of gender equality. We looked at how well these six criteria were applied.

There was significant variation across evaluations as to which of these criteria were addressed and to what extent. This is reflected in our discussion of each criterion.

This chapter first presents an overview of the findings from our assessment against the 15 criteria and then moves on to more detailed analysis for each criterion.

[24] The robustness of the evaluators' numerical ratings against the standard M&E criterion was not considered in this review; however, the review does consider the evaluation's assessment of the adequacy and use of the initiative's M&E system under general evaluation quality standard criterion 5.

Overview of the findings

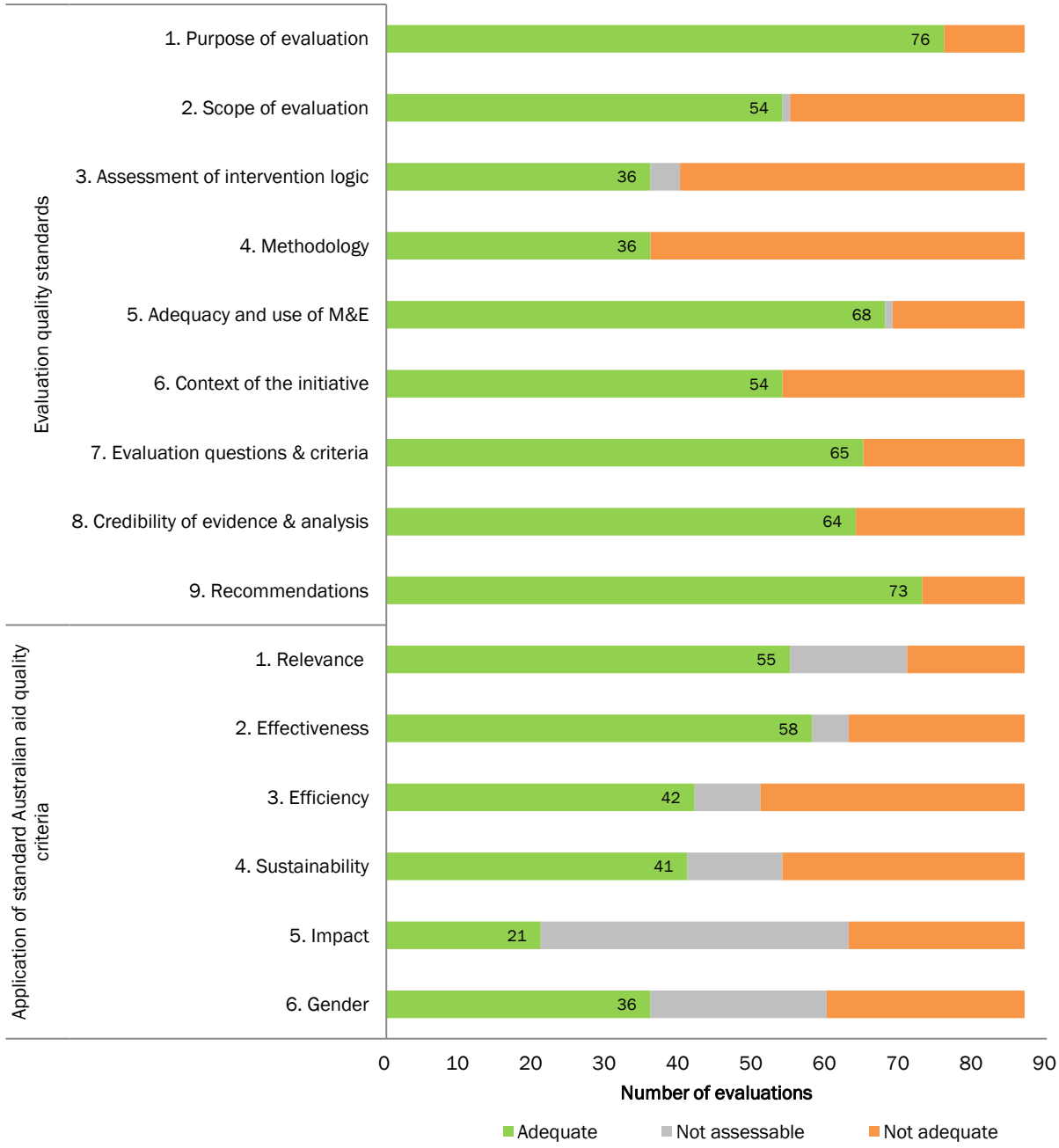
In reviewing the quality of evaluation designs, we found that most evaluations had a clear purpose. Just over half either did not assess the logic of the intervention or provided a weak assessment and did not adequately justify the evaluation methodology used. While the evaluation guidelines at the time did not explicitly require either an assessment of logic or a justification of evaluation methodology (a description of the methodology was required), these would have strengthened the quality of the final evaluation reports. Greater focus on evaluation design, particularly through the routine requirement for an evaluation plan, may help improve overall evaluation quality. Specialist support to commissioning areas in the early stages of an evaluation is also likely to help address these evaluation design weaknesses.

For the key criteria relating to the quality of the final evaluation report, we found that quality was adequate or better in most cases. The credibility of the evidence and analysis was adequate or better in 74 per cent of evaluation reports, showing that independent initiative-level evaluations do generally provide a credible source of evidence for the effectiveness of the Australian aid program. Other areas of strength were the quality of the evaluations' assessments of the initiatives' monitoring and evaluation (M&E) systems and the quality of the recommendations. Many evaluations did not, however, adequately consider the influence of context on initiative performance.

When we looked at the quality of the evaluations' assessments against the standard Australian aid quality criteria we found that most assessments of relevance and effectiveness were adequate or better. However, only about half of the assessments against the other criteria were adequate. Where numerical ratings were provided we found that across five criteria (excluding impact) about two-thirds were robust. Only half of the evaluations attempted to assess the impact of the aid initiative, and fewer than half of those assessments were adequate.

Figure 3 presents a summary of our assessment results for the 15 quality criteria. Figure 8 in Annex 4 presents detailed results.

Figure 3 Summary of 87 evaluations against 15 quality criteria[25]



[25] 'Not assessable' was recorded where there was insufficient information in the evaluation report to make a judgment. Where the report did not address a particular criterion because it was not (or did not appear to be) included in the scope of the evaluation, no assessment was made. There were two other cases where a criterion could not be assessed for other reasons: one evaluation (INI691) covered 13 different trust funds and assessing intervention logic for all was unfeasible; the terms of reference for the other evaluation (INI426) were not available and the scope of the evaluation could not be assessed based on the information in the evaluation report alone.

Box 4 Good practice examples

To highlight instances of model practice and support ongoing learning and improvement within the aid program, our review identified several examples of good practice evaluation documents:[1]

- Independent progress review of Partnership for Knowledge-Based Poverty Reduction (INJ244)—terms of reference, evaluation plan, and evaluation report
- Independent completion report for Timor–Leste Asian Development Bank Infrastructure Project Management/Infrastructure Technical Assistance (INH497)—terms of reference
- Independent evaluation of AusAID's support to rural WASH in Timor-Leste through the Rural Water Supply and Sanitation Program (ING002)—evaluation plan
- Independent completion review of the Civil Society Water, Sanitation and Hygiene Fund (INI592)—evaluation report
- Independent review of two remote service delivery and community development programs in Papua New Guinea (INH843 and INJ153)—evaluation report

These examples are referred to throughout sections 4.1 and 4.2 and discussed in detail in Annex 6.

4.1 Quality of evaluation designs

Four of the general evaluation quality criteria relate to the quality of the underlying evaluation designs. While documents such as terms of reference and evaluation plans typically provide detailed information on the evaluation design, these documents were not available in all cases, so our assessments were based primarily on evidence from the evaluation reports themselves (although many reports included these documents as annexes and we did consider them where they were available).

Clarity of purpose of evaluation

We found the clarity of the purpose of the evaluation to be adequate or better in 87 per cent of evaluation reports. This indicates that, on the whole, the overall focus of evaluations of Australian aid initiatives is being clearly articulated. The purpose was often couched in the language of 'accountability', 'identifying lessons', 'driving improvement' and 'informing future decisions'.

It is worth noting, however, that fewer than half of these evaluations explicitly detailed the primary audience for the report, and that these frequently referred only to 'AusAID' without additional details. This is important as it indicates the up-front planning that has gone into how the evaluation findings are going to be used and who is to take them forward.

For terms of reference that provide a clear outline of the audience for an evaluation, see the good-practice example from the evaluation of Timor–Leste Asian Development Bank Infrastructure Project Management/Infrastructure Technical Assistance (INH497) in Annex 6.

[1] To be selected as a good-practice example, an evaluation document had to meet the following minimum standards. For the terms of reference, criterion 1 (evaluation purpose) was assessed as good quality; for the evaluation plans, criterion 4 (evaluation methodology) was assessed as good quality; for the evaluation reports, all criteria were assessed as adequate quality or better. We sought to identify at least one evaluation that illustrated good practice across the evaluation report, terms of reference and evaluation plan. This required us to be flexible in our application of the evaluation report criteria.

The true purpose of the evaluation should be made clear from the beginning and should inform the evaluation design. During interviews three consultants mentioned cases where decisions about the future of an initiative had been made before an evaluation began but had not been communicated to the evaluation team. While it is sometimes unavoidable to undertake an evaluation after key programming decisions have been made, there may still be good opportunities for learning.

Scope of evaluation

We found that 62 per cent of evaluations adequately considered the scope of the evaluation, including whether the scope matches the evaluation resources and whether the roles of the evaluation team members and management are set out.

It is worth noting that, of the 41 cases (47 per cent of evaluations) where AusAID officers were included in the evaluation team, in half of the cases their role was not clearly defined.

Assessment of intervention logic

Our review found that there was adequate assessment of the underlying logic of the intervention in only 42 per cent of evaluations. Typically evaluations failed to assess the intervention logic or theory of change, or the clarity of the initiative's objectives. Even evaluations rated highly against other criteria mostly assessed the intervention logic in an implicit way only, and there was rarely an explicit section devoted to assessment of the intervention logic.

While the terms of reference for many evaluations, especially independent progress reports, did not ask for a critique of the intervention logic, and the evaluation guidelines at the time did not explicitly require an assessment of the intervention logic, it was surprising that more evaluation-commissioning areas and evaluators were not more inquisitive on this point. Lack of assessment of an intervention's logic does not preclude gathering robust evidence, but assessment of logic provides a structure and coherence to an evaluation process that significantly strengthens the analysis. The conclusion is that the majority of evaluations are failing to really test the underlying logic and objectives of the intervention—which in most cases should be a key rationale for commissioning an external evaluation.

Evaluation methodology

A common weakness was that, while evaluation reports described the methods used, most offered no justification as to their suitability for the purpose of the evaluation and/or their possible strengths and weaknesses. We found the evaluation methodology to be adequately justified and the use of sources to be adequate in only 40 per cent of all evaluation reports. In arriving at this assessment, we considered justification of the design of the evaluation and whether a clear explanation of the techniques for data collection and analysis was provided. In addition, we looked for evidence of triangulation between types of evidence and, where applicable, an appropriate sampling strategy, discussion of the limitations of the methodology, and consideration of ethical and cultural issues.

The treatment of methodology in the majority of reports tended to be formulaic. Many simply listed interviews and focus groups as the specified methodology without reflecting on their suitability compared to other possible methods or providing details on exactly how they would be used to gather the necessary data. Most reports provided little or no explanation of the sampling strategy used. While we acknowledge that this is not an explicit requirement of the evaluation standards or guidelines, it does make it difficult to understand why certain people were interviewed and not others. While general limitations like lack of time or access were mentioned, reports rarely discussed the

limitations of specific methods, or whether the methodology had to be changed as the evaluation proceeded.

We expected the methodology sections to at least touch on issues of attribution and contribution, including the extent to which changes or results can be attributed to the Australian aid program, and on counterfactual issues like what might have happened without a particular initiative. However, these issues were addressed very rarely in relation to methodology and infrequently in other parts of the evaluation reports.

Most reports lacked an adequate description of the ethical issues faced by an evaluation and how they were addressed. Even when such issues were covered, there was usually only a brief discussion of confidentiality and little discussion of other issues.

4.2 Quality of evaluation reports

The remaining five general evaluation quality criteria relate to the quality of the final evaluation reports themselves.

Credibility of evidence and analysis

A strength identified through the review is that the credibility of evidence and analysis was adequate or better in 74 per cent of evaluation reports. This is an important criterion as it tests the underlying quality of the evidence base and the robustness of the analysis. It covers seven sub-criteria:

- findings flow logically from the data
- any gaps or limitations in the data are explained
- assumptions are made explicit
- conclusions, recommendations and lessons are substantiated
- the position of the author is clear
- issues of attribution and contribution to results are discussed
- alternative views/factors are explored to explain the observed results.[2]

The importance of this criterion has been confirmed by testing the statistical association between each of the 15 quality criteria. This indicated that 'credibility of evidence and analysis' is the criterion most strongly associated with the other criteria. It is therefore the best predictor of quality—that is, if an evaluation received a good rating for this criterion then it was also likely to have good ratings for many of the other criteria (see Annex 4, Figure 10, for more details of this analysis). We have therefore used the 'credibility of evidence and analysis' criterion throughout this report as a proxy indictor of overall evaluation report quality.

[2] These seven sub-criteria were equally weighted. An evaluation had to be found adequate or better for the majority of these sub-criteria to be rated adequate overall for credibility of evidence and analysis. A limitation of this method is that an evaluation report may fail to address important sub-criteria yet still be assessed as adequate. (It is worth noting that this means that credibility of evidence and analysis could in some cases be assessed as adequate even while the methodology was assessed as inadequate. This was possible because, even if some aspects of the evaluation methodology were unsatisfactory, the logic and flow of argument from findings to conclusions in the evaluation, and hence its credibility, could still be satisfactory.) The credibility or robustness of any ratings awarded by the evaluator was not considered under this criterion; this is examined separately in section 4.3.

Our analysis of this criterion found that a particular strength was the logical and clear line of evidence running through the evaluation report, leading to well-substantiated conclusions and recommendations. This in turn often led to a clear, unambiguous judgment being made. The majority of evaluation reports also clearly explained the gaps and limitations in the data.

Two good examples of evaluation reports with these characteristics are discussed in Annex 6: the review of two remote service delivery and community development programs in Papua New Guinea (INH843 and INJ153) and the evaluation of the Civil Society Water, Sanitation and Hygiene Fund (INI592).

The evaluations tended to be weaker in providing analysis of the extent to which the initiative contributed to the outcomes observed and in considering alternative explanations for the outcomes. In the majority of evaluations, the contribution of the initiative to an observed change was assumed and was not interrogated in depth. This is a notable gap in the evaluation reports and is reflected in the fact that, of those we judged to be adequate quality or better, more than half (33 out of 64) were rated merely adequate rather than good or very high quality. This finding is also consistent with the finding of a recent meta-evaluation of 340 US Agency for International Development evaluation reports[3] that that only 10 per cent discussed other possible causes in addition to USAID interventions that might be contributing to results. One of the explanations for the inattention in Australian aid evaluations to issues of attribution and contribution to results could be that 59 per cent of the reviewed evaluations were independent progress reports that took place midway through initiative implementation. A detailed analysis of contribution is less likely in such reports as they tend to be more focused on process and on documenting activities and outputs achieved than on outcomes and impacts.

Adequacy and use of monitoring and evaluation systems

Our review found that the majority of evaluations (79 per cent) had adequate analysis of the quality of M&E systems and use of M&E data. It should be emphasised that this high score is not a comment on the quality of these systems but rather refers to the quality of the evaluations' often extensive, and usually critical, coverage of this issue. Issues around the quality of initiative M&E systems are discussed briefly in section 5.1.

An example of a good assessment of the quality of the initiative M&E system is in the evaluation of Partnership for Knowledge-Based Poverty Reduction (INJ244) discussed in Annex 6 (see pp. 9–10, 19 and 40 of the evaluation report).

Evaluation questions and criteria

Our review found that 75 per cent of the evaluation reports adequately addressed the standard quality criteria or the evaluation questions / major issues detailed in the terms of reference.

A good example where each evaluation question and sub-question is answered systematically is the review of the Livelihoods and Food Security Trust Fund in Myanmar (INJ135).

[3] M. Hageboeck, M. Frumkin & S. Monschein (2013). *Meta-Evaluation of quality and coverage of USAID evaluations 2009–2012*, Management Systems International under subcontract to DevTech Systems, Inc. Cited in USAID (2013), *Discussion note on complexity-aware monitoring*.

Recommendations

The recommendations were found to be adequate or better in 84 per cent of evaluation reports. The recommendations in those reports were clear and relevant.

> A particularly good example of clear recommendations is in the evaluation report for the Timor–Leste Asian Development Bank Infrastructure Project Management/Infrastructure Technical Assistance initiative (INH497) (pp. 34–35).

However, our review indicated room for improvement in how actionable the recommendations were and the extent to which they considered resource implications. While it may not always be possible or appropriate for an evaluation to consider the resource implications of its recommendations, ensuring that recommendations are actionable and targeted is a key factor in ensuring that the evaluation will be useful.

Context of the initiative

One area of weakness in the evaluation reports was the quality of the contextual analysis. Only 62 per cent of evaluation reports were found to be adequate in this area. Those that were inadequate tended to describe the context of the initiative but not analyse its influence on performance.

4.3 Quality of the evaluations' assessments against the standard Australian aid quality criteria

This section considers the quality of the evaluations' assessments against six of the standard Australian aid quality criteria, where those criteria were addressed. The section then also considers the robustness of the numerical ratings against these criteria in the evaluations (fewer than half of the total) where they were provided. As discussed in Chapter 2, at the time when the evaluations covered in this review were completed, evaluations were expected to consider and provide a rating against the six standard Australian aid quality criteria outlined below.

Relevance

The assessments of initiative relevance were adequate or better in 77 per cent of the 71 evaluations that addressed this criterion. In these cases, the evaluations analysed and provided a solid judgment on the extent to which the initiative aligned with the goals that Australia shares with its development partners in a given context. In line with the standards widely applied when assessing the relevance of Australian aid initiatives, this was our minimum threshold for an assessment of adequate.

However, it could be argued that this reflects too narrow a view of relevance. Some evaluations did consider other elements of relevance—as suggested in departmental guidance—including partner government perspectives, alignment with Australian Government strategies, context, and the appropriateness of the aid delivery modality.

> An example of a strong response to the relevance criterion is in the evaluation of Partnership for Knowledge-Based Poverty Reduction (INJ244) discussed in Annex 6 (see pp. 21–27 of the evaluation report).

Interview responses suggested that the exact meaning of relevance is not consistently well understood by evaluators. Both Australian Government aid managers and consultants raised questions about how issues of relevance were addressed in these reports.

Effectiveness

The assessments of aid initiative effectiveness were adequate or better in 71 per cent of the 82 evaluations that addressed this criterion. These evaluations provided an evidence-based assessment of progress towards expected outcomes. They offered a clear definition of what effectiveness should mean in the context of the initiative and what level of results can be expected given the stage of implementation when the evaluation occurs. The better assessments for this criterion presented a good balance of qualitative and quantitative evidence and analysed each component of the initiative in a systematic way. They provided a clear and often compelling link between the evidence presented and analysed and the conclusions drawn. Sources for the evidence were given and, where there were doubts over data quality or data gaps, these were discussed.

An example of a strong assessment of effectiveness is in the evaluation of the Civil Society Water, Sanitation and Hygiene Fund (INI592) discussed in Annex 6 (see Chapter 3 in the evaluation report). Another good practice example is the partner-led mid-term review of the State- and Peace-Building Fund (ING948), which showed very clear links between project results and strategy (pp. 19–26). The evaluation of the Zimbabwe NGO Food and Water Initiative (INJ189) provided detailed evidence in appendices 5 to 8, and discussed risk management in several places (e.g. pp. 18, 47 and 51).

The 29 per cent of evaluation reports rated as inadequate in their assessment of effectiveness tended to fall short due to their failure to address the extent to which the initiative contributed to the outcomes observed, and to consider alternative explanations for the outcomes. Being able to link decisions, actions and deliverables as contributing to an emerging outcome is an important aspect of assessing effectiveness; however, this was not usually mentioned. This reflects our findings on credibility of evidence and analysis (discussed in section 4.2).

Within the assessments of initiative effectiveness, an area that was not addressed well was the treatment of risk as it affects the achievement of outcomes. Risk was specifically discussed in only 14 cases. Risk management issues are often not relevant for ex-post or completion reviews but we would have expected them to have a higher profile in independent progress reports given the potential impact of different risks on the effectiveness of an initiative.

Efficiency

The assessments of initiative efficiency were adequate in only 54 per cent of the 78 evaluations that addressed this criterion. This low score reflects the fact that the majority of evaluations took a very narrow view of efficiency, limiting their focus to issues such as management processes, staffing and disbursement delays. Very few reports substantively engaged with broader issues of value for money such as cost-efficiency or cost-effectiveness. These findings reflect those of the 2011 Bazeley study.

Even fewer evaluations discussed the costs of achieving outputs/outcomes through alternative means. While a full value for money analysis may be beyond the scope of most evaluations, a meaningful discussion around alternative strategies that could have been used to achieve an outcome and their relative cost should be possible.

Sustainability

The assessments of sustainability (i.e. of the benefits that will endure after Australia's contribution has ceased) were adequate in only 55 per cent of the 74 evaluations that assessed this criterion. The most common weakness was that evaluations often failed to explore what sustainability means in the context of the initiative and to unpack what the key dimensions of sustainability are for the likely long-term success of the initiative. A number of evaluations approached sustainability only from the angle

of resourcing and failed to explore other aspects such as levels of local ownership, and integration with existing institutions and systems.

Impact

Initiative impact was not well covered in the evaluation reports reviewed: it was assessed in only 45 of the 87 evaluations (52 per cent)—the lowest coverage of any criterion. Where impact was not covered, this was almost always because it was not included in the terms of reference.[4] Of the 45 evaluations that did consider impact, fewer than half (21) had adequate or better impact assessments and there were only 12 examples of good-quality impact assessment.

The low coverage and quality of impact assessments is to some extent related to the fact that 59 per cent of the evaluations were independent progress reports, while long-term impact is typically not apparent until well after the completion of an aid initiative. The quality of impact assessments—where included—was better in independent completion reports (57 per cent were adequate) than in independent progress reports (39 per cent were adequate). This is understandable given that the purpose of independent progress reports is rarely to measure impact but is rather to assess implementation progress.

Under DFAT's current evaluation planning and resourcing models for independent evaluation of aid initiatives, an evaluation is required only once during the life of an initiative rather than after its completion and the resources required to properly evaluate the long-term impact of an initiative are typically not readily available. Furthermore, it can be difficult to assess impact in the absence of strong performance data, including baseline data. The uneven coverage and quality of impact assessments suggests the need to review the guidelines on when impact should be evaluated and what aspects should be considered.

Gender equality

We found that assessment of the extent to which the initiative promotes gender equality (i.e. develops and implements appropriate and effective strategies to advance gender equality and promote the empowerment of women and girls) was adequate in only 57 per cent of the 63 evaluations where this was assessed. In the subset of evaluations where the assessment was good (24 of 63 evaluation reports), gender equality was usually seen as central to the evaluation and plenty of evidence was presented. For the 43 per cent of gender assessments found to be inadequate, discussion of gender was often superficial, with only a brief paragraph devoted to the topic. Some reports simply mentioned the number of women reached by a particular initiative and did not discuss issues of gender equality and empowerment.

Robustness of numerical ratings

As well as looking at the evaluations' written assessments against DFAT's standard quality criteria for Australian aid, our review also checked the robustness of the numerical ratings against the same criteria, where such ratings were provided. The inclusion of numerical ratings in evaluation reports was expected under guidelines in effect in 2012; however, fewer than half of the 87 evaluations actually provided numerical ratings (as discussed in section 3.2). For those evaluations that did provide numerical ratings, we checked the robustness of the rating by comparing the evidence

[4] There were also three cases where the terms of reference were not available, so it was not clear whether the evaluators were asked to assess initiative impact.

presented in the report against the rating. If the evidence substantiated the rating awarded by the evaluation team then it was marked 'robust'. If the evidence was insufficient then it was marked 'not robust'. If the evidence appeared to merit a different rating, either higher or lower, it was marked accordingly. The results are presented in Figure 4.

Overall, 66 per cent of the ratings for relevance, effectiveness, efficiency, sustainability, and gender equality were robust. (The robustness of ratings for M&E was not checked. Impact is discussed separately below.) The most robust ratings were for relevance, with 74 per cent of ratings found to be robust. Ratings against each of the other four criteria were robust in at least 60 per cent of cases. In most cases where ratings were not robust it was usually because they were too high based on the evidence presented. This indicates the need for evaluators to ensure that they are not overrating initiative performance in light of the actual evidence to hand.

Only nine of the 87 evaluations provided a numerical rating for impact, and we assessed only four of these as robust. While the sample size is too small to make any generalisations, this seems consistent with our findings on the poor coverage and quality of impact assessments.

Figure 4 Robustness of evaluations' numerical ratings against standard Australian aid quality criteria

	Relevance	Effectiveness	Efficiency	Impact	Sustainability	Gender
Number of the 87 evaluations with a rating	38 (44%)	38 (44%)	38 (44%)	9 (10%)	38 (44%)	32 (37%)
Number of robust ratings	28 (74%)	24 (63%)	23 (60%)	4 (44%)	25 (66%)	21 (66%)
Total number of not robust ratings	10	14	15	5	13	11
a. Too low	0	2	2	0	0	1
b. Too high	9	11	10	4	11	8
c. Otherwise not robust[5]	1	1	3	1	2	2

[5] Cases where it was not possible to assess whether the rating was too low or too high because the evidence presented was too weak to justify any rating.

5 Factors influencing evaluation quality and utility

This chapter identifies underlying factors contributing to or inhibiting evaluation quality and utility. The findings are presented thematically and, where possible, we have triangulated between quantitative and qualitative evidence and analysis. We present only the findings that emerged most clearly.

The quantitative analysis examined the relationship between our quality assessments against the 15 quality criteria and a range of variables or possible explanatory factors. We focused in particular on the 'credibility of evidence and analysis' criterion as the one summary measure of evaluation quality. This was based on the principle that this criterion is not only the most important of the quality criteria but also has the highest level of correlation with the other criteria (as discussed in section 4.2). Our analysis was limited in many cases by small or incomplete data sets.

Qualitative evidence was collected during interviews with 14 consultants who had responsibility for writing the evaluation reports, 10 Australian Government aid initiative managers and three Australian Government senior executives.[6] The two former groups of interviewees were selected on the basis of their association with evaluations assessed across several criteria as either adequate quality or inadequate quality. The key purpose of the interviews was to gain a deeper understanding of the factors that enable or inhibit evaluation quality.

While we did not set out to make evaluation utility a focus of this review, the evidence from interviews raised significant issues around evaluation utility and suggested that evaluation quality and utility are closely linked.

5.1 Characteristics of the aid initiative

We found some variation in evaluation quality depending on the value of the aid initiative. Most notably a lower proportion of evaluations for very large initiatives were found to be of adequate quality. There was also some variation in evaluation quality across sectors.

Initiative value

On average, more resources tended to be applied to evaluating larger initiatives. Up to a point, higher initiative value corresponded with higher evaluation quality. Of the 63 evaluations of initiatives valued at less than $35 million (the average value of the evaluated initiatives), 71 per cent were found to have credible evidence and analysis. Of the 15 evaluations for larger initiatives valued between $35 million and $99.9 million, 87 per cent were found to be credible. However, of the nine evaluations for very large initiatives valued at $100 million or more, only six, or 67 per cent, were found to be credible. While the small number makes generalisation problematic, this does suggest

[6] See Annex 2 for details of our selection method.

that greater attention to assuring the quality of evaluations for very large value initiatives may be needed.

Sector

We analysed the quality of the evaluations by sector, using the 'credibility of evidence and analysis' criterion as a proxy for overall evaluation quality.[7]

The infrastructure, humanitarian response, education, and general development support sectors had a higher than average proportion of credible evaluations.

The food security and rural development sector and the human rights sector had a lower than average proportion of evaluations assessed as credible. In particular, the human rights sector had only five out of 10 evaluations assessed as credible. We were not able to identify clear reasons for weaker evaluation quality in these two sectors, although both had lower than average value initiatives. It may be worth focusing more support and quality assurance in these sectors.

Other sectors were either comparable to the overall average or did not have enough evaluations to make a comparison. See Annex 4, Figure 11 for more detailed findings of our analysis.

Quality of initiative monitoring and evaluation system

Another important initiative characteristic likely to have an impact on evaluation quality is the quality of the initiative's monitoring and evaluation (M&E) system, as this affects the quality of the performance data to which evaluators have access. As discussed in section 3.2, of the 35 evaluations that provided a numerical rating for the quality of the initiative's M&E system, only 18 (51 per cent) rated that system as adequate or better. While we would have liked to conduct further correlation analysis to determine the nature and strength of the relationship between M&E system quality and evaluation quality, we did not have enough numerical data on the former. However, evaluation findings regarding the quality of initiative M&E systems are synthesised in our *Learning from Australian aid operational evaluations* report, and a separate ODE study is proposed for 2014 on the quality of initiative M&E systems.

5.2 Evaluation purpose and type

We found significant variation in quality between independent progress reports and independent completion reports. Overall, the quality of evaluations managed by the former AusAID was at least as good as that of partner-led joint evaluations, and this was achieved with fewer resources. Some interviewees suggested that utility could be improved by reviewing the balance between the numbers of operational and strategic evaluations produced.

---

7 The first step of the analysis confirmed that the overall level of association between findings for 'credibility of evidence and analysis' and sector was significant in statistical terms—i.e. some of the variability in one can be accounted for by the other. Association was tested using a chi-squared statistic that measures the probability of ratings scores for different categories of data (in this case sectors) being drawn from the same population. The chi-squared test for independence indicated a significant probability that there is a statistical difference between sectors (chi-squared = 80, p = 0.04 where significance is indicated by a value of 0.05 or less).

Independent progress reports and independent completion reports

A higher proportion of independent progress reports than independent completion reports were of adequate quality. Using the credibility of evidence and analysis criterion as a proxy for overall evaluation report quality, we found that 76 per cent of the 51 independent progress reports were adequate or better; most of these were just adequate.

Only 68 per cent of the 34 independent completion reports were adequate or better. Moreover, the quality of independent completion reports was more variable: compared to independent progress reports, a larger proportion of independent completion reports were good or better but also a larger proportion were poor or worse, as shown in Figure 5.

Figure 5 Credibility of evidence and analysis in independent progress reports and independent completion reports

The reasons for this difference are not clear but could be connected to evaluation resourcing. As discussed in section 3.2, independent progress reports had a higher average level of resourcing.[8] The difference could also be connected to the purpose of the evaluation. A larger proportion (90 per cent) of independent progress reports than independent completion reports (82 per cent) were found to have a clear purpose. The purpose of an evaluation completed at the close of an initiative may sometimes be less clear to aid managers and evaluators in terms of providing opportunities to inform critical programming decisions. This could, in turn, have an impact on evaluation quality. The perceived value of evaluations is discussed further in section 5.6.

Partner-led joint evaluations

Our review found that, overall, the quality of evaluations managed by the former AusAID was at least as good as that of partner-led joint evaluations and that this was generally achieved with fewer resources.[9] Each had different areas of comparative strength and weakness. A slightly higher proportion of the 72 AusAID-led evaluations (75 per cent) had credible evidence and analysis,

8 As discussed in section 5.3, we did not have enough evaluation budget data to determine correlations with evaluation quality.

9 Length of evaluation, number of fieldwork days and evaluation team size were used as proxy indicators for evaluation resourcing in the absence of reliable evaluation cost data for the partner-led joint evaluations. See section 3.2.

compared to 67 per cent of the 15 partner-led evaluations. A significantly higher proportion of evaluations led by the former AusAID had adequate recommendations, assessments of initiative M&E systems and assessments of gender equality. On the other hand, a significantly higher proportion of partner-led joint evaluations had adequate assessments of the context and of the impact of the initiative. Quality was roughly comparable for all other criteria (see Figure 11 in Annex 4 for detailed findings). While the small sample of only 15 partner-led joint evaluations makes us wary of making too much of these differences, this is a positive finding for the Australian aid program, especially when differences in resourcing levels are taken into account.

Balance between operational and strategic evaluations

Some interviewees suggested that utility could be improved by reviewing the balance between the numbers of operational initiative-level evaluations and strategic evaluations produced. A few interviewees questioned the current balance between the two, and suggested that the Australian aid program should aim for fewer, more strategic evaluations. One consultant suggested that the Australian aid program should consider 'Commissioning more evaluations of the delivery strategy in one country: too many evaluations focus on the level of activities and outputs, and too few really confront the key strategic issues'. A senior aid executive noted that the Australian aid program 'does struggle to point to solid evaluations that show our broader contribution [other than single interventions]'. These views suggest that the current focus on evaluating initiative-level outcomes may be diverting attention and resources away from evaluating how well the Australian aid program is achieving its higher-level outcomes.

5.3 Evaluation resourcing

As expected, our findings suggest that evaluation quality is influenced by the level of resourcing provided. We did not have enough reliable evaluation cost data to look for correlations with evaluation quality; however, we were able to use proxy indicators. As discussed in section 5.1, we were able to determine a correlation between initiative value and evaluation cost and between initiative value and evaluation quality (with the notable exception of very high value initiatives). We were also able to use the duration of the evaluations as a proxy indicator for the level of evaluation resourcing and then look for correlations with evaluation quality.

Duration of the evaluation

We found a positive correlation between evaluation duration (as measured by the total evaluation length and by the number of fieldwork days[10]) and overall evaluation quality (as measured by the findings for the credibility of evidence and analysis criterion)—see Figure 6. Evaluations found to be good or very high quality averaged, respectively, 61 and 79 days in overall length and 18 and 25 days of fieldwork. Evaluations found to be poor or very poor quality were much shorter, averaging 27 days overall length and nine days of fieldwork.

[10] Of the two measures, the correlation was marginally stronger between evaluation quality and number of fieldwork days (correlation coefficient of 0.36), than between quality and overall length of the evaluation (correlation coefficient of 0.30). The number of total days recorded for an evaluation may be a less reliable measure, as there are sometimes delays or breaks in the evaluation work that can extend the total amount of time used. Other aspects of time use in evaluations, such as analysis or write-up time, could not be analysed as insufficient data was available. Positive correlations (mostly weaker) were also found for the majority of the other evaluation quality criteria.

Figure 6 Assessed 'credibility of evidence and analysis' and evaluation length

Measure of evaluation length	Assessed credibility of evidence and analysis				
	Poor or very poor quality	Less than adequate quality	Adequate quality	Good quality	Very high quality
Average number of evaluation days	27	37	34	61	79
Average number of fieldwork days	9	13	14	18	25

The positive correlation between evaluation quality and number of fieldwork days was confirmed by our interviewees, 11 of whom emphasised the need to have sufficient time in country for evaluations. Interviewed consultants stressed the importance of adequate lead-in and preparation time and the challenges posed by rushed mobilisation and inadequate time for report writing. Their views were supported by aid program staff interviewees who noted that evaluations need to be planned well in advance (usually around six months) so that the best or most appropriate consultants can be contracted and there is sufficient time to prepare and set up meetings with stakeholders in advance of fieldwork.

5.4 Evaluation team

We found some variation in evaluation quality depending on the size of the evaluation team. Other related factors contributing to evaluation quality included the expertise in the evaluation team, the quality of the relationship between the evaluation team and Australian Government aid program staff, and the involvement of aid program staff in an evaluation.

Evaluation team size

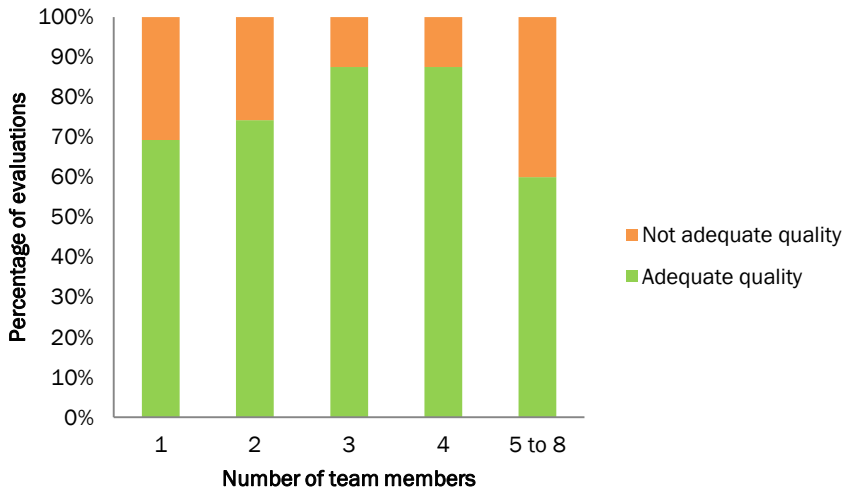
Our analysis shows that teams of three or four produced a higher proportion of adequate evaluations than smaller teams of one or two, or larger teams of five or more. We analysed evaluation quality (as measured by findings against the credibility of evidence and analysis criterion) by team size (not including any Australian government aid program staff who may have participated in the evaluation[11])—see Figure 7. For teams of three or four, 88 per cent of evaluations were assessed as adequate for credibility of evidence and analysis. By comparison, 73 per cent of the evaluations produced by single consultants or teams of two were assessed as adequate and only six of the 10 evaluations produced by teams of five or more (60 per cent) were assessed as adequate.[12]

It should be noted that team size on its own is not likely to be a simple determinant of quality. We conducted additional analysis to see whether the optimal team size might vary according to the size of the initiative being evaluated, but we were not able to determine a clear pattern.

11 Information on their level of participation was not consistently available.

12 This difference was found to be statistically significant. It was significant at the 90% level using a chi-squared test (i.e. there is only a 1 in 10 possibility that this difference was due to chance).

Figure 7 Assessed 'credibility of evidence and analysis' and evaluation team size



Evaluation team expertise

A key factor contributing to evaluation quality was the expertise of the consultants, particularly the team leader. This factor was mentioned by almost one-third (eight out of 27) of interviewees, including three Australian Government aid managers. Team expertise combines a number of competencies—above all a strong technical knowledge of different evaluation methodologies; knowledge of how to lead an evaluation and manage both international and local consultants; strong diplomatic and interpersonal skills; expertise in collecting, analysing, and presenting data; and ability to write credible reports in a tight timeframe. It also includes relevant expertise in the particular sector and, to a lesser extent, some understanding of the country or regional context. Two Australian Government aid managers (including a senior manager) noted the crucial importance of identifying team leaders and team members with these qualities.

The fact that it can be difficult to find all these qualities in one or two people may partially explain the finding (discussed above) that evaluation teams of three or four members produced a higher proportion of credible evaluations.

Relationship between evaluation teams and Australian aid program staff

A quarter of interviewed consultants and Australian Government aid managers emphasised the quality of the relationship between evaluation teams and commissioning aid managers as another key factor contributing to both evaluation utility and quality. In the best cases, evaluations have benefited from a high degree of mutual respect and mutual learning, and this positive relationship has increased the utility of evaluations. Aid initiative managers indicated that they are more likely to implement the recommendations if they have full confidence in both the evaluation team and the overall process.

Strong relationships can also benefit the technical quality of the evaluations. For example, when consultants are able to fine tune the terms of reference in consultation with aid initiative managers at the outset, they are likely to have a more nuanced understanding of the key objectives of an evaluation. There were positive examples where the consultants and aid managers were able to build their relationship to the extent that the consultants were invited to do follow-up work. Several consultants also stressed that, where their contact with an evaluated initiative had continued after

completion of the evaluation report and they had been able to see how their reports were utilised—and, in particular, which of their recommendations were implemented—they valued the continuing relationship and felt that this knowledge could help them improve the quality of future evaluations.

But interviews also highlighted problematic cases in which trust had broken down due to personality clashes or the consultants' perception that the Australian Government aid managers were not being open about decisions that had already been made about the future funding of an initiative being evaluated. (Presumably this would not be a problem if it were made clear from the outset that the evaluation was being undertaken for other reasons.)

Involvement of Australian aid program staff in evaluations

A key issue raised by the majority of consultants as an enabling factor, and in some cases as a constraint, was the extent to which Australian Government aid program staff were involved in an evaluation. This includes initiative and program managers and also advisers such as gender specialists or sector specialists. In the best cases they were seen as having an essential role in facilitating the whole evaluation process (particularly in terms of access to information and key stakeholders). Their involvement can also assist consultants to frame their recommendations and ensure that evaluation findings are taken seriously. One consultant interviewed said 'the lower the level of staff involvement, the higher the risk that it won't hit the target and be taken seriously'.

However, of the five consultants interviewed who had Australian Government aid program staff in their evaluation teams, four found the involvement of those staff problematic. In particular they suggested that the presence of Australian Government aid program staff in meetings with key stakeholders threatened the independence of the evaluation. One consultant commented about a staff member that 'the inhibition that her presence created was incalculable'. A particular difficulty mentioned was that interviewees receiving Australian Government funding were unlikely to express criticism in the presence of Australian Government staff, and that this could potentially compromise the evaluation findings. The common theme that emerged from our interviews is the need for clarity about the staff member's role in the evaluation. For example, one consultant said: 'There were two [Australian aid program] staff on the team, and their role as evaluators was not clear. There is a one-page note [from the Australian aid program] on the role of staff on independent evaluations[13] but they had not seen this.'

As one would expect, the views of Australian Government aid program staff on the involvement of staff in evaluations were more mixed. Some saw it as a valuable learning opportunity for staff. However, one who had been directly involved in an evaluation had experienced some uncertainty about whether he should be accountable to AusAID or to the evaluation team leader.

The evidence suggests that some involvement by Australian Government aid program staff can be helpful but that it is important to clearly define and actively manage a staff member's role in an evaluation from the outset, including ensuring that team leaders feel empowered to decide which meetings or interviews staff attend.

---

13 Guideline: Independence in AusAID evaluations (internal document).

5.5 Evaluation design

The evidence suggests that quality assurance of evaluation designs may help improve overall evaluation quality. Both our analysis of evaluation reports and our interviews emphasised the importance of well-thought-out terms of reference and evaluation plans. While we did not directly assess the quality of these documents (because they were not consistently available), our assessment of the quality of the evaluation reports revealed a need to improve the focus on quality at the early stages of an evaluation.

Terms of reference

As discussed in section 4.1, over half of the evaluation reports either did not test the underlying logic of the intervention or had a weak assessment.

Consultants interviewed noted cases where their terms of reference had either been far too limiting or, conversely, far too ambitious for the resources available. In two cases consultants said they appreciated being able to refine terms of reference in consultation with Australian aid program staff. One said that before starting an evaluation '[Australian aid program] staff should encourage consultants to question and if necessary change the terms of reference'. (As discussed below, this negotiation can alternatively take place as part of the development of an evaluation plan.)

Our interviews also suggest that it is useful for the terms of reference to attempt to prioritise the key criteria at the outset, so that the evaluators can prioritise their time accordingly.

Evaluation plans

As discussed in section 4.1, our quality review also found that over half of the evaluation reports reflected weak or inappropriate evaluation methodology, or did not adequately justify the evaluation methodology. This suggests a need for greater focus on improving the quality of evaluation plans.

Evaluation plans are recommended under the current guidelines but are not mandatory. Before an evaluation commences, a detailed evaluation plan is developed by the evaluator based on the terms of reference and negotiations with the DFAT evaluation commissioning area. The evaluation plan should elaborate on the terms of reference and provide detail on the methodologies to be followed.

However, two interviewed consultants stressed that when both consultants and aid managers feel under time pressure the evaluation plan might be seen as another bureaucratic requirement rather than a real opportunity for dialogue about evaluation methodology. This view is supported by the identified weakness of evaluation methodology and specifically by a common failure to adapt evaluation methods to meet the requirements of a specific evaluation. As one team leader said:

> *The process of preparing for an evaluation and doing a plan… has become bureaucratised and worthless. It's become a cut and paste activity. A smart evaluator will look at what previous evaluators have done and cut and paste either what they or others have done. If they are particularly good, they might customise the list for a specific evaluation, but it's a 'tick the box' activity.*

The evidence suggests a need for greater emphasis on the importance of developing strong evaluation plans as the basis for high quality evaluation reports. This includes allowing sufficient resourcing (consultancy days) for the evaluator to develop the plan, including negotiating with the DFAT evaluation commissioning area. It may also indicate a need to improve support for, or quality assurance of, evaluation plans.

5.6 Organisational factors

This section considers a number of aspects of the organisational culture of the former AusAID that are likely to have had an impact on evaluation quality and utility. It is based on evidence gathered through interviews and contextual analysis. Some of the issues raised are already recognised within the department and may be at least partially addressed through planned changes to simplify the aid management system.

Evaluation capacity

Interviews revealed a perception by both staff and consultants that there is often a trade-off between the quantity of evaluations produced and evaluation quality. In particular, concerns were raised that pressure on Australian aid program staff to deliver more evaluations has meant that specialised evaluation skills are stretched. Less experienced staff and staff without specialist evaluation skills are being required to commission and manage evaluations. Some interviewees (including senior managers) noted that sometimes, due to other work pressures, the management of evaluations was given to relatively junior staff who did not feel well equipped for this task. Without adequate technical support, evaluation quality can be negatively affected.

One senior manager expressed these concerns about the Australian aid program's overall evaluation capacity:

> *The rapid growth in [the Australian aid program] is stretching us. There is good ambition in the agenda, but we don't necessarily have the internal skills and systems to support this ambition… The key is how can we do better, by doing slightly less. Everyone wants to do more, but there is a real risk that this will sap the ability of posts to get value out of this. The more we mandate, the more we insist on templates, the less learning we'll see. An evaluator comes in with external eyes and provides lots of learning. We need to make evaluations more focused, and make it easier for posts to do these.*

While acknowledging that there is now greater technical support for evaluations commissioned by posts than in the past, some staff interviewed noted that the high turnover of staff in posts meant that there was a continuing need for technical support. Some junior staff interviewed requested more basic evaluation training as part of their induction, as well as more hands-on tailored support for specific evaluations.

Evaluation culture

Comments from many interviewed staff and consultants suggested that the value of evaluation as a tool for performance management and learning is not being fully realised. Several interviewed aid managers expressed the view that evaluations are frequently undertaken only to comply with requirements, and that learning from evaluations could be improved both for individual aid managers and collectively across the aid program. There was a perception that, at present, the aid program may be commissioning more evaluations than it can effectively learn from. The following quotations from interviews show this tension:

AusAID is one of the most evaluation heavy agencies, and we commission independent evaluation and approaches as a 'tick-box' exercise. Reports are shelved. No one reads them. I think there is far too much emphasis on evaluation at the expense of monitoring. If it weren't such an 'evaluation heavy' culture, when it does come to evaluation there wouldn't be such a gap in information.

(Initiative manager)

AusAID culture is more a 'tick box and put on shelf' attitude to evaluations. AusAID needs to be more organic, and feed lessons learned into the agency better. We don't have systems to capture lessons or feed them into programming. So we should pull together evaluations and then target users.

(Initiative manager)

If the managers see the merit, they can get an awful lot out of an evaluation. But if they delegate the management of evaluations to junior staff who do not understand the scope of what is being achieved then the result is more likely a 'box ticking' exercise. We therefore have a product [evaluation reports] that is highly variable.

(Senior manager)

Our evidence and analysis suggests that current organisational incentives risk promoting a compliance-driven approach to evaluation at the expense of developing a culture of learning from evaluations.

The senior management attitude to evaluation can be critical to determining whether evaluation is seen as 'just another bureaucratic requirement' for initiative managers or as a critical opportunity both to learn from experience and to improve outcomes. The senior managers we interviewed were clear about the centrality of independent evaluation and the need for learning from evaluations, and were able to pinpoint particular programs that they felt had made particularly good use of evaluations in planning future strategies and initiatives. As can be seen by the quotations above, not all staff appear to be aware of these views of senior management.

Perceived pressure to show positive results

Interviews also revealed another tension that could be having an impact on both evaluation quality and utility: on one hand the need for robust evaluations to inform performance management and learning, and on the other hand a perceived pressure to show positive results for the Australian aid program, coupled with demand for greater transparency.

There was a surprising level of agreement between consultants and staff about the pressures they felt from a drive within the aid program to show positive results for the Australian taxpayer. Two consultants reflected on how these pressures might impact on the quality and independence of evaluation reports:

It's inevitable that consultants 'listen to the corporate view' and adjust findings to it.

On the one hand we want to do the best job possible, but we also know we have to deliver a service… we have to understand what AusAID needs from the evaluation… there is lots of tension between what donor wants/expects and also the need to educate the donor about program realities.

Senior aid executives also referred to the tension felt in the Australian aid program (as in most other donors and large international NGOs) between the twin pressures of greater transparency and the need to show positive results. However our interviews with senior managers suggested that, as one would expect in any large and dynamic organisation, there is no single 'corporate view'; indeed one senior manager saw the major challenge as identifying evaluators who do have the expertise and authority to be truly independent and to challenge conventional wisdom.

While it is difficult to find strong evidence on the extent to which the perceived pressure to show positive results for the Australian aid program has had an effect on evaluation quality or utility, it is clear that such organisational incentives should be taken into account in any efforts to improve the aid management system.

5.7 Access to completed evaluations

As well as being used by evaluation commissioning areas for performance management and learning purposes relating to a single initiative or program, evaluations can have utility for a wider audience. Evaluations can provide lessons for managers of similar aid initiatives in other programs; the findings of multiple evaluations can be synthesised to identify lessons for improving aid management (as we have done in the *Learning from Australian aid operational evaluations* report); and ratings against the standard Australian aid quality criteria—where provided—can be analysed to identify patterns and trends in performance. Ease of access to completed evaluations and evaluation data is critical to whether evaluations are used in these ways, and to the development of a departmental culture of learning from evaluations.

Our experience in assembling the evaluations for this review revealed significant difficulties in accessing evaluation data. Completed evaluations of aid initiatives are stored internally in AidWorks, the department's aid management system. However, there is insufficient standardisation in the way they are stored: uploaded independent evaluations are not always recorded as such, and many documents labelled as 'independent evaluations' are actually other types of documents. Compiling the list of evaluations for this review required extensive manual checking. Furthermore, numerical ratings against the standard aid quality criteria cannot be exported from AidWorks for analysis; they have to be manually transcribed from each evaluation report. These factors limit the use of evaluations by DFAT staff outside the area that commissioned an evaluation. These factors (compounded by flexibility in the timing of evaluations during the life of an initiative) also make tracking compliance with evaluation requirements difficult and resource intensive.

There are similar issues with access to published evaluations of Australian aid initiatives by external stakeholders. As mentioned in section 3.1, about half of the operational evaluations completed in 2012 were published on the Australian aid website. However, access to these evaluations is not straightforward; they are not all accessible from a central location in the website and are not always clearly marked as evaluations, so they may not turn up in searches.

Recent changes aimed at promoting better evaluation pipeline planning at the program level will increase the visibility of evaluations and thus, to some degree, the ability of interested parties to access and learn from them. In 2013 a requirement was introduced that all annual Aid Program Performance Reports include an annex listing planned, current and recently completed evaluations. Continued strengthening of program-level evaluation planning has been emphasised in the context of simplification of the aid management system. As one interviewed initiative manager suggested:

Each post needs to have an annual schedule of evaluations and link up programs more. AusAID has so many evaluations, it needs to plan them far better in advance, with annual schedules for each post, and it must plan out more how learning at post (country) level can be synthesised.

While program-level evaluation planning is a positive step, further improvements to facilitate easier access both internally and externally to evaluations and evaluation data would also support the development of a culture of learning from Australian aid evaluations.

6 Conclusions and recommendations

6.1 Assessed quality of operational evaluations

Coverage of the aid program

Compliance with mandatory operational evaluation requirements was satisfactory. The initiatives evaluated were diverse in terms of value, sector and geographic region, and can be considered broadly representative of the overall Australian aid program.

Evaluation designs

A greater focus on improvement at the evaluation design phase (evaluation terms of reference and evaluation plans) may help strengthen overall evaluation quality. We found that most evaluations had a clear purpose. However, just over half had weak or no assessment of the underlying logic of the intervention, or did not adequately justify the evaluation methodology used. This may reflect the absence of any specific guidelines requiring either an assessment of the underlying logic or the justification of methodology. It may also relate to the capacity of non-specialist staff to commission high-quality evaluations. This finding suggests a need to focus greater support and quality assurance efforts at this early stage. Ensuring sufficient lead time and resourcing (consultancy days) for the evaluator to develop and negotiate a strong evaluation plan may also help.

Evaluation reports

We found the overall credibility of the evidence and analysis in the evaluation reports to be satisfactory in 74 per cent of cases. Most other aspects of the quality of the evaluation reports were also satisfactory in the majority of cases. The quality of the evaluations' assessments of the initiative's monitoring and evaluation (M&E) systems and the quality of recommendations were good overall. However, report quality could be improved by a stronger focus on analysis of the extent to which initiatives contribute to the outcomes observed and on the influence of context on initiative performance.

Evaluations' assessments against the standard Australian aid quality criteria

We found that most of the evaluations' assessments of the relevance and effectiveness of the aid initiative were of adequate quality or better, which suggests that operational evaluations are generally a robust source of evidence about the effectiveness of the Australian aid program.

However, only about half of the assessments against other standard Australian aid criteria—efficiency, sustainability and gender equality—were adequate.

Most notably, we found that consideration of initiative impact was frequently omitted from evaluation terms of reference and was not assessed. In cases where it was assessed, the proportion of assessments found to be adequate was low. This is understandable given that impact is typically not

apparent until well after the completion of an initiative, whereas the evaluations we reviewed (except for the two ex-post evaluations) were conducted during the implementation of the initiative, often mid-implementation. The inability of most evaluations to examine impact suggests the need for further thinking within the department as to how best to capture and consider aid impact.

Numerical ratings against the standard Australian aid quality criteria

Numerical ratings against the standard Australian aid quality criteria were provided in fewer than half of the evaluation reports. Where they were provided, the vast majority of initiatives were rated as adequate or better for relevance and effectiveness. The proportion rated adequate for efficiency, gender equality and particularly M&E was significantly lower. We found that about two-thirds of numerical ratings (excluding impact) were robust, suggesting a need to ensure that evaluators are not overrating initiative performance in light of the actual evidence at hand. The low coverage and questionable robustness of ratings raises questions about their purpose and utility.

Evaluation resourcing

As is to be expected, our findings suggest that evaluation quality is influenced by the level of resourcing provided. We did not have enough reliable evaluation cost data to look for direct correlations with evaluation quality but we were able to use proxy indicators.

On average, more resources tended to be applied to evaluating larger initiatives. Up to a point, higher initiative value corresponded with higher evaluation quality. However, it was concerning to find a lower than average proportion (six out of nine, or 67 per cent) of evaluations for very large initiatives ($100 million or greater value) to be adequate quality. This may in part reflect the complexity of many very large initiatives.

We also found a positive correlation between evaluation quality and evaluation duration. Evaluations of good or very high quality averaged, respectively, 61 and 79 (calendar) days in overall length, and 18 and 25 (calendar) days of fieldwork. This suggests that evaluation commissioning areas should allow sufficient time for completion of evaluations, including adequate fieldwork time.

Type of evaluation

We found a higher proportion of independent progress reports than independent completion reports to be of adequate quality. Moreover, the quality of independent completion reports was more variable: a higher proportion was inadequate but also a higher proportion was good quality. The reasons for this difference are not clear but could be connected to the higher average level of resourcing for independent progress reports. It could also be connected to the level of clarity about the purpose of the evaluation: sometimes the purpose is less clear at the close of an initiative in terms of providing opportunities to inform critical programming decisions.

Overall the quality of evaluations managed by the former AusAID was at least as good as that of partner-led joint evaluations and this was generally achieved with fewer resources.

Evaluation team

We found that evaluation teams of three or four members (not including any Australian aid program staff) produced a higher proportion of adequate quality evaluations than teams of one or two or teams of five or more. This suggests that areas commissioning evaluations should ensure that there are enough members in the evaluation team to cover the range of expertise required for the particular

evaluation. Several interviewees highlighted the quality of the relationship between evaluation teams and commissioning aid managers as a key factor contributing to both the quality and the utility of evaluations.

The evidence suggested that involving Australian aid program staff in an evaluation can have numerous benefits, including better understanding of the operational context, better uptake of recommendations, smoother facilitation of the evaluation process (including access to information and key stakeholders), and developing the evaluation capacity of Australian aid program staff. However, this involvement needs to be actively managed to avoid the risk of compromising the independence of the evaluation. The role of any DFAT staff should be clearly defined from the outset, and evaluation team leaders should feel empowered to decide which meetings or interviews staff attend.

Quality of initiative monitoring and evaluation systems

The quality of an initiative's M&E system also has a significant impact on evaluation quality, as it affects the quality of the primary data to which evaluators have access. We found the quality of initiative M&E systems to be weak in a large proportion of cases. We did not have enough data to determine the precise nature and strength of the relationship between the quality of initiative M&E systems and the quality of evaluations; however, around half of the 35 evaluations that provided a numerical rating for the quality of the M&E system rated it as inadequate. Furthermore, our *Learning from Australian aid operational evaluations* report found that:

- there needs to be more investment in developing clearer intervention logic and robust monitoring arrangements at the initiative design phase
- M&E systems need to maintain a stronger focus on outcomes rather than outputs
- M&E data needs to be kept simple and accessible so that it can be used as the basis for decision-making.

A separate ODE study is proposed for 2014 on the quality of initiative M&E systems.

6.2 Departmental capacity to manage the volume of evaluations

A central feature of the department's evaluation policy is mandatory initiative-level evaluation. As a consequence, a large cohort of program staff are required to commission and manage evaluations as part of their normal program management duties. In recent years there have been significant efforts across the department to boost the capacity of non-specialist staff to help deliver high-quality evaluations. Nevertheless realistic expectations need to be maintained as to the degree of evaluation expertise and knowledge these staff can or should obtain. Several interviewees indicated that, given the volume of evaluations undertaken, the evaluation capacity within the department is particularly stretched. Some interviewees also suggested that evaluation numbers are beyond the optimum for performance management and learning and do little to assist the cross-program use of evaluations and the development of a department-wide culture of learning from evaluations.

In early 2012 the department's evaluation policy was revised, reducing the number of mandatory operational evaluations by approximately half, to only one during the lifetime of each aid initiative. Further changes planned for mid-2014 will raise the financial threshold for aid initiatives requiring mandatory evaluation and will reduce evaluation numbers by a further 42 per cent. The overall reduction in evaluation numbers is accompanied by the introduction of country and regional program evaluation plans to help improve the allocation of resources and skilled staff. Structural changes have

also been made with the responsibility for monitoring and reporting on operational evaluations shifting to ODE and marking a move towards consolidation of aid evaluation expertise within the department.

6.3 Support arrangements for operational evaluations

With the shift of the operational evaluation oversight function from the Program Effectiveness and Performance Division (PEPD) to ODE in February 2014, it is timely to consider the types of support arrangements likely to best promote improvement in the quality of operational evaluations across the department. There is demand from program areas for more hands-on tailored support for particular evaluations (e.g. assistance with defining terms of reference, identifying consultants suitable for a particular evaluation, negotiating evaluation plans, and reviewing evaluation reports) in addition to the support provided by program-based performance and quality managers.

There is a potential role for ODE in quality assurance of evaluation documents. In recent years there has been little quality assurance of evaluation documents beyond peer reviews of draft evaluation reports, when it is too late to make changes to the scope of the evaluation or the methodology used. Our review suggests that it would be appropriate to focus support and quality assurance efforts on the evaluation design phase (evaluation terms of reference and/or evaluation plans) and on the evaluation of high-value investments.

6.4 Access to completed evaluations

Our team experienced significant difficulties in accessing evaluation reports in AidWorks, the department's aid management system, and in accessing the evaluations published externally. Such difficulties limit the use of evaluations outside commissioning areas. Ease of access to evaluation reports is critical to the development of a culture of learning from evaluations. Storage issues also make tracking compliance with evaluation requirements difficult and resource intensive.

While recent changes aimed at promoting better program-level evaluation planning are a positive step, further improvements to facilitate easier access to evaluations both internally and externally are also needed. While improving knowledge management is a large-scale and long-term undertaking, it is worth considering whether there are immediate options for improving the accessibility of evaluation data.

6.5 Evaluation purpose

Evaluations are commissioned for various purposes, including to drive improvement, to inform future programming decisions and to provide accountability. This is reflected in guidelines for operational evaluations, which allow a high degree of flexibility in evaluation timing and resourcing.

Our review found that evaluation commissioning areas do exercise a high degree of flexibility in the timing of operational evaluations. Evaluations can be undertaken during initiative implementation (independent progress report) or at the close of an initiative (independent completion report). We found that more independent progress reports are undertaken than independent completion reports; that independent progress reports had a higher average level of resourcing than independent completion reports; and that a higher proportion of independent progress reports had a clear purpose and were of satisfactory overall quality. This may indicate that independent progress reports are more

useful to aid managers and evaluators in terms of providing information and evidence to inform critical programming decisions such as whether to extend an investment.

Flexibility in scoping

However, departmental guidelines, while not mandatory, do not encourage flexibility in scoping. They set out expectations that operational evaluations will assess initiative performance against a set of standard Australian aid quality criteria (relevance, effectiveness, efficiency, sustainability, impact, and gender equality).

Our review found evidence to suggest that this can lead to an evaluation scope that is too ambitious to be realistic or appropriate, and that on occasion this may negatively affect evaluation quality and utility. While the assessments against the key aid quality criteria of relevance and effectiveness were generally strong, about half of the assessments against the other criteria were weak and were sometimes completed in a perfunctory manner (especially gender equality). Furthermore, several interviewees suggested that they or their colleagues view operational evaluations as 'box-ticking' exercises without clear utility. This suggests a need for clearer messaging that operational evaluations should be scoped to meet the particular information needs of program areas.

Evaluating the impact of the aid initiative

Half of the evaluations did not attempt to assess the long-term impact of the aid initiative. Where impact was assessed, the quality of the majority of those assessments was weak. This raises questions about the appropriateness of including the assessment of impact as standard in operational evaluations, given that the impact of an aid initiative is difficult to assess until well after its completion and typically relies on a robust monitoring and evaluation system during its lifetime. Consideration may be needed as to whether impact should remain a standard quality criterion for operational evaluations.

Rigorous assessments of end-of-program outcomes and of impact remain a high priority to inform learning and account for the results of public spending on aid. Special arrangements for commissioning and resourcing evaluations specifically designed to look at the long-term impact of aid initiatives should be considered, particularly for high-value investments and/or those that offer broader learning opportunities. Such evaluations serve a distinct and important purpose. They do, however, need to be properly resourced. To ensure that such complex evaluations are completed to a high standard, they would need to be managed or supported by staff with high-level specialised evaluation skills. There may be a role for ODE to jointly manage some impact evaluations with program areas.

6.6 Lessons for DFAT evaluation commissioning areas to maximise evaluation quality and utility

The evidence from our quality review of operational evaluations highlights the following lessons for DFAT evaluation commissioning areas. They are valuable lessons not only for the operational-level officers who commission evaluations, but also for the senior executives responsible for resource planning and for making programming decisions based on the performance information generated by the evaluations.

These lessons should be read in conjunction with departmental evaluation guidelines, particularly the DFAT Aid Monitoring and Evaluation Standards, which articulate expectations of the quality expected

from a range of M&E products. The relevant standards relating to independent evaluations are included at Annex 5. The standards had not been formally adopted at the time the evaluations in this review were undertaken, but they were integrated into evaluation guidance in 2012. The Standards provide a useful resource for evaluation commissioning areas.

Our review also identified several examples of good practice evaluation documents, which are discussed in Annex 6.

Box 5: Lessons for DFAT evaluation commissioning areas

Start evaluation planning well in advance

1. Consider the timing of an evaluation when you are developing initiative monitoring and evaluation arrangements. Plan the timing of the evaluation so that it will be most useful for program management purposes. Use program-level evaluation planning to help with the allocation of resources and skilled staff.
2. Start planning the evaluation six months ahead of planned commencement. Adequate time is needed to develop good-quality terms of reference (seeking support from performance and quality managers if needed), contract the most suitable consultants, engage with the consultants in their evaluation planning and schedule fieldwork to allow access to key stakeholders. Options may be limited if there is insufficient lead time.

Focus on developing strong terms of reference as the basis for a good-quality evaluation

3. Using the aid quality criteria as a starting point, develop key evaluation questions that address the most critical issues and management decisions related to the initiative. Prioritise these questions to ensure a focus on the things that really matter. Consider including assessment of the intervention logic or theory of change. (If the intervention logic is not clearly articulated in the initiative design or implementation documents then one of the first evaluation tasks should be to reconstruct the intervention logic.)
4. Allocate sufficient time for the evaluation. This should match the scope of the evaluation but for a good quality evaluation would typically be two to three months from when the consultants commence to when the evaluation report is finalised. In particular, allow enough days for the consultants to develop a strong evaluation plan with a methodology that is appropriate to the evaluation questions, and for fieldwork.
5. Consider the skills required within the evaluation team, and the number of evaluation team members needed to cover this range of skills. Evaluation teams should consist of people with technical evaluation expertise and strong interpersonal skills, in addition to sectoral expertise and, to a lesser extent, knowledge of the country or regional context.
6. Be clear about the roles of any DFAT staff involved in the evaluation.

Continue to actively engage with the evaluation team during the evaluation

7. Invest time and effort in building strong relationships with the evaluation team.
8. Debate contentious issues, but respect the independence of the evaluation. Allow the team leader to exercise judgment on participation of staff in meetings or interviews.

6.7 Recommendations

Acknowledging the need to improve the evidence base for effective aid programming and the principles of simplicity, proportionality and value for money, this review recommends that:

Recommendations

Recommendation 1

DFAT should review arrangements (including responsibility and resourcing) for the following evaluation functions:

- evaluation planning at program level, including prioritisation and resourcing of evaluations
- support by dedicated evaluation staff for non-specialist evaluation managers, particularly for developing evaluation terms of reference and/or evaluation plans and for evaluation of high-value investments.

Recommendation 2

DFAT should make it explicit that the purpose of the evaluation guides the approach to that evaluation. Specifically:

- operational evaluations should remain flexible in timing, with their scope and methodology purposefully designed to meet the specific information needs of program areas
- consideration should be given to commissioning impact evaluations, especially of high-value investments and/or those that offer broader learning opportunities.

Recommendation 3

DFAT should monitor implementation of the policy requirement to publish all independent operational evaluations and should improve their public accessibility.

Annex 1: Terms of Reference

[**These Terms of Reference were finalised in March 2013, before the absorption of the former AusAID into the Department of Foreign Affairs and Trade.**

Subsequent to the finalisation of these Terms of Reference, the proposed review was split into two separate reports: *Quality of Australian aid operational evaluations* and *Learning from Australian aid operational evaluations*.]

The Office of Development Effectiveness (ODE) and the Program Effectiveness and Performance Division (PEPD) of AusAID will jointly manage a review of the quality and synthesise the findings of independent evaluations commissioned by AusAID of aid initiatives. This topic (synthesis of AusAID evaluations) is included in ODE's forward work plan which was endorsed by the Independent Evaluation Committee (IEC) and approved by the Development Effectiveness Steering Committee.

1 Background

Independent Evaluation at AusAID is undertaken at several levels and managed by different areas. ODE undertakes evaluations of broad strategic relevance in line with its evaluation policy and three year rolling work program. Thematic areas commission sector evaluations (such as the Mid-Term Review of the Development for All Strategy, 2012) and geographic areas also commission evaluations (such as the Review of the PNG–Australia Development Cooperation Treaty, 2010). However, the bulk of independent evaluations are undertaken at initiative level. In accordance with AusAID's Performance Management and Evaluation Policy (PMEP), every monitored initiative[14] is required to undertake an independent evaluation at least once over its life, at the time and for the purpose most useful for program management. (This replaces an earlier policy that distinguished between Independent Progress Reports (IPRs) and Independent Completion Reports (ICRs)). The purposes of these independent evaluations are:

- **Management:** Independent evaluations help managers to understand what is working, what is not and why, and feed directly into improved management by informing initiative quality at implementation assessments and annual program performance reports, the ODE synthesis of evaluations and quality assurance report and the *Annual Review of Aid Effectiveness*.
- **Accountability:** Independent evaluations are a key source of information on the effectiveness of the aid program to key stakeholders, such as the Australian public, partner governments, implementing partners and the communities that AusAID works with.

[14] A 'monitored' initiative is where the expected Australian Government funding over the life of the initiative is greater than $3 million, *or* the value is less than $3 million but the initiative is significant to country or corporate strategies or key relationships with other development partners, including other government agencies.

› **Learning**: Independent evaluations provide important information about what does or does not work in a particular context and why. This information may inform country and thematic strategies, design of new activities, management of existing ones, and provide learning to the global community.

For initiatives that are co-financed with other donors or implemented through partners, AusAID encourages joint or partner-led evaluations to be undertaken to share learning across all partners, and to avoid over-burdening implementing partners and beneficiaries with multiple evaluation processes. These evaluations are regarded as meeting AusAID's requirement to undertake an independent evaluation, and are expected to be published on the AusAID website.

With the exception of the Australian Centre for International Agricultural Research (ACIAR), for most other Australian Government departments delivering the aid program, little data is available on evaluations conducted. In cases where those departments are funded through AusAID's budget, they are required to comply with AusAID's PMEP. Departments who directly appropriate aid funding follow their own performance management processes. In 2013 the Development Effectiveness Steering Committee (DESC) endorsed Whole of Government Uniform Standards for the aid program, including a standard on performance management under which all Australian Government departments must conduct an independent evaluation at least once over the life of every aid project.[15]

There are three drivers for this review of independent, initiative-level evaluations:

1.1 The need for effectiveness reporting on the Australian aid program to draw on a body of credible evidence

An Effective Aid Program for Australia states that Australia's approach will be based on 'concrete evidence of what works best on the ground to produce results'. Evaluations are central to this aid effectiveness and results agenda, in driving ongoing learning which informs the direction, design and management of the aid program. Independent evaluations also play an important accountability role in AusAID's performance management systems. They complement annual performance management processes which are based on self-assessment, and provide an independent perspective of the quality and results achieved through the Australian aid program.

In line with recommendations in the *Independent review of aid effectiveness* to strengthen initiative and program evaluation in AusAID, the government committed in Effective Aid to producing a smaller number of high-quality evaluations. Under the Transparency Charter, it is expected that evaluations will be published.

1.2 The need to improve the quality of the aid program's independent evaluations

Previous meta-evaluations of AusAID's independent evaluations have found issues with compliance with agency evaluation requirements, and the quality of initiative-level evaluations.

In response to issues identified in ODE's 2007 *Review of AusAID's approach to evaluation*, an Evaluation Review Panel was established by ODE in September 2008 to improve the quality of evaluations in AusAID, and also to build the capacity of AusAID officers to recognise the quality or otherwise of independent evaluations. A blind technical review process was used where consultants were asked to review and provide a technical rating for draft evaluation reports. Over 70 evaluations

[15] ODE will work with AusAID's Whole of Government Branch as it works to apply uniform standards to ODA managed by other government agencies.

underwent technical review via this process between 2008 and 2010. An evaluation of the process[16] found that the technical review could be improved, but should be continued. Nonetheless, the Technical Review Panel was discontinued in 2011. Since then, quality assurance of evaluation findings/reports has been through peer review, rather than the previous two-step technical review plus peer review system.

The March 2011 *Study of independent completion reports and other evaluation documents* (the Bazeley study)[17] raised concerns regarding compliance with agency evaluation requirements, and with the quality of evaluations. For example, Bazeley found that evaluations were undertaken primarily for accountability purposes and not for learning or management, poor underlying data from M&E systems, and the average time allowed (23 days) was minimal given evaluation expectations.

Since April 2011 there have been no further meta-evaluations of the quality of AusAID evaluations or the evaluation process.

1.3 The increasing importance of evaluation across the Australian Public Service

The Department of Finance and Deregulation is overseeing a process of renewing evaluation processes across the Australian Public Service as part of the Commonwealth Financial Accountability Review (CFAR). ODE is a member of the inter-departmental committee advising on this issue. As the new financial accountability framework comes into place across the Australian Public Service, it will be important for the aid program and AusAID, including ODE, to maintain a high-quality and systematic program of evaluations. This review of operational evaluations will help to position the role of evaluations across the aid program.

2 Scope

The 2013 review of operational evaluations will consider all independent evaluations of initiatives completed by AusAID and/or partners (where that evaluation is used for AusAID internal purposes) in the 2012 calendar year.

Under the current Performance Management and Evaluation Policy, AusAID expects approximately 111 initiative evaluations to be undertaken each year.[18] A recent stocktake of evaluations conducted for the 12 months ending October 2012[19] identified 103 independent evaluations having been undertaken during that period, excluding ODE, thematic and geographic-based evaluations. A relatively small number have been published.

Future reviews of operational evaluations may move to a financial year reporting period to align with other corporate reporting processes. Future reviews may also look at other types of evaluations (e.g. thematic evaluations, ODE evaluations, evaluations by other government departments). For the

---

16 Patricia Rogers, Meta-evaluation of AusAID's technical review process, RMIT University, April 2011.

17 Commissioned in support of the Independent review of aid effectiveness. This study reviewed evaluations from a four-year period from July 2006 to June 2010.

18 In the 2011–12 financial year there were 588 monitored initiatives with an average duration of 5.3 years. Every initiative is required to undertake an evaluation at least once over its life. So, assuming an even split of evaluations per year, approximately 110–111 evaluations would be expected in 2012.

19 Since the Bazeley study, which covered the 2006–2010 financial years, no financial year stocktakes of completed evaluations have been undertaken.

purposes of this review, the basic characteristics of these other evaluations will be briefly considered by way of context.

All evaluations will be included in the quality review component of the review of operational evaluations. A selection of these evaluations will be used in the synthesis component.

3 Objectives

In line with the purposes of evaluation in AusAID, and the quality issues highlighted in the meta-evaluations outlined above, this review of operational evaluations has two objectives:

- to promote good quality independent evaluations (including appropriate coverage)
- to inform the Minister, public, partners and aid program staff of overarching lessons emerging from The findings of independent evaluations.

The findings from the review of operational evaluations will also provide input for ODE's 2014 synthesis of evaluations and quality assurance review and the 2014 *Annual Review of Aid Effectiveness*. It is anticipated that the review will become a regular product and this will be reflected in the agency's PMEP.

4 Focus questions

The review of operational evaluations will seek to answer the following questions.

Quality review

1. What are the basic characteristics of different levels of independent evaluation in the aid program and the history and nature of independent evaluation at the initiative level?
2. To what degree do independent evaluations[20] provide a credible source of evidence for the effectiveness of the Australian aid program?
3. What are the major strengths and weaknesses of independent evaluations conducted for AusAID?
4. What are the factors that contribute to their quality?
5. What actions should be taken to improve the quality and/or coverage of independent evaluations?

Synthesis

1. What are the main lessons for the aid program emerging from the findings of independent evaluations?
2. Are there any trends or patterns regarding the effectiveness, relevance, sustainability or other characteristics of evaluated initiatives?[21]

---

20 'Independent evaluations' is hereafter used in these terms of reference to mean initiative-level independent evaluations.

21 This question was dropped. The question had assumed that a high number of the evaluation reports under review would provide numerical ratings for quality. However, the quality review revealed that only 40 per cent of evaluations provided any numerical ratings, so the question was no longer relevant.

5 Approach

The review of operational evaluations will be conducted by a small team of consultants, and jointly managed by ODE and PEPD (Quality Performance and Results Branch), with input from a reference group comprised of AusAID senior management/advisers. The review will be overseen by the IEC.

Preparatory phase

PEPD will look at the population of monitorable initiatives and conduct a compliance check against the evaluation requirements set out in the PMEP. PEPD will collate a list of all independent evaluations which have been completed (i.e. the date of the final evaluation report) in the 2012 calendar year.

To identify any patterns in coverage or compliance, PEPD/ODE will analyse AidWorks data to compare the characteristics of the initiatives for which independent evaluations have been completed (including, for example, stage of implementation, value, location, sector, implementing partner, modality) with those for which they haven't been completed (including those for which exemptions were granted). This analysis will feed into the quality review.

In addition, the number and characteristics of thematically-based evaluations completed in the 2012 calendar year (i.e. the date of the final evaluation report) will be identified. This data will inform the quality review; however, these evaluations will not themselves be quality reviewed.

A review plan providing details on the agreed methodology and how the review will be implemented will be prepared, and endorsement sought from the IEC. This review plan will be revised before the commencement of part 2: synthesis.

Part 1: Quality review

Part 1 of the review of operational evaluations is a meta-evaluation to assess the credibility and quality of evaluation reports, including the major strengths and weaknesses of independent evaluations and contributing factors. It will also identify actions that should be taken to improve the quality and/or coverage of independent evaluations.

Key activities will include:

- assessing each evaluation against the OECD–DAC evaluation criteria and a selection of the 2013 AusAID M&E Standards. A clear method and pro forma will be developed to assess the credibility and quality of each evaluation's assessments against each of the criteria[22]
- conducting analysis to determine whether there are any correlations between the quality of evaluations and the characteristics of the initiative (for example, value, location, sector, implementing partner, modality), or between the quality of evaluations and the characteristics of the evaluation (for example, stage of implementation, length of evaluation, time taken for evaluation fieldwork and reporting, focus on project or sector issues, degree of country-specific analysis)
- conducting interviews with evaluators and program managers from a sample of evaluations to help identify the factors contributing to stronger or weaker evaluations, primarily focusing on using an 'appreciative inquiry' approach.

[22] For example, the ALNAP pro forma: www.alnap.org/pool/files/QualityProforma05.pdf.

During the quality review, approximately six (6) examples of 'good practice' evaluation products (terms of reference, evaluation plans, and evaluation reports) will be identified. These examples illustrating 'good practices' identified during the quality review of the evaluation population will be discussed in an annex to the final report, with a focus on providing learning to AusAID staff. At least two good quality examples of each type of evaluation product will be sought.

Following an approach yet to be decided, feedback on the assessment of individual evaluations will be provided to initiative/program managers.

In the case that the consultants in the review team have been involved in undertaking an evaluation that is subject to quality review, or in designing or implementing that initiative, they will be recused from conducting the quality assessment for that evaluation to avoid a conflict of interest. The AusAID management team will identify a suitable substitute quality reviewer for the evaluation in question, and this will be acknowledged in the report.

Part 2: Synthesis

Part 2 of the review of operational evaluations will be a synthesis of insightful and useful lessons from a selection of evaluations with particular characteristics (for example, sector, location, implementing partner, modality).

The following approach will be taken:

- Using a methodology to be developed in consultation with the consultants, the synthesis focus and sample will be determined, drawing on the analysis of the characteristics of evaluations during the quality review and discussion regarding possible focus areas with the reference group and relevant program/thematic areas. Through this process, more specific evaluation questions will be developed for the synthesis and the final sample selected on the basis of these questions.
- It is anticipated that a maximum of 60 evaluation reports will be included in the synthesis.
- Findings from the individual evaluations which fall within the synthesis sample will be analysed and synthesised. This may include interviews with key specialist staff and/or seeking to compare the synthesised findings with findings from other international evidence sources (particularly if there is clear contradiction or correlation) in order to explore particular issues in more depth.
- The findings of the synthesis will be tested through peer review with the reference group and other subject matter specialists, country specialists or modality specialists (depending on the focus areas covered).

6 Outputs

A **review plan** will provide details on the methodology to be used and how the review will be implemented. The review plan will be prepared by the review team (consultants), and will be endorsed by the IEC prior to the quality review commencing. This review plan will be reviewed after part 1: quality review has been completed.

The key output will be a **final report** presenting the findings of the quality review and the synthesis. This report will summarise the evidence collected, present analysis and findings, and make recommendations where appropriate. The report will be approximately 30–35 pages in length (plus 4-page executive summary), and will include a quality review section of a maximum of 15 pages and a synthesis report of maximum 15 pages. The report will include a context section that describes the characteristics of different types of independent evaluation conducted at different levels of the aid

program (ODE, thematic, geographically based, other government department and initiative level) before focusing in more detail on the history and nature of initiative-level independent evaluations since about 2006. The report will be prepared by the review team (consultants) in two separate parts, and will be reviewed by ODE, PEPD, the reference group, peer reviewers (part 2: synthesis only) and the IEC.

Approximately six (6) **examples of 'good practice' evaluation products** will be identified and discussed in an annex to the final report. If possible, the examples chosen will be from initiatives with diverse characteristics (e.g. sectors, geography, implementing partner). The examples should include at least two (2) examples of each of a range of good practice products (e.g. terms of reference, evaluation plans, evaluation reports).

Detailed records of all evidence collected or analysis undertaken (including records of the quality assessments, interview notes, spreadsheets containing raw data) will be retained by ODE and PEPD for possible future analysis, but will not be included in the report.

7 Roles and responsibilities

An **AusAID management team** comprising one Director and one manager from ODE and one Director and one manager from PEPD will collaboratively manage the review of operational evaluations. During the preparatory phase, this team will collate the independent evaluations for review, and undertake analysis of AidWorks data, subject to the availability of resources. The team will agree on methodology and comment on reports from the consultants. If the team cannot agree through a collaborative approach, issues may be taken to the reference group for resolution. However, in the case of any issue arising that cannot be resolved collaboratively, the Assistant Director General ODE will make a determination.

A **review team** of up to four consultants with skills in evaluation, analysis and report writing will prepare the review plan, undertake the quality analysis and the synthesis using a methodology agreed with the AusAID management team, and prepare the draft report.

A small **reference group** comprised of AusAID senior management/advisers will be consulted at key decision points.

The **Independent Evaluation Committee (IEC)** will provide technical oversight of the review of operational evaluations. The final report will be made public as an ODE product.

8 Timeframes

Date	Activity	Primary responsibility
Preparatory phase		
April - May 2013	Collate independent evaluations and conduct preliminary analysis	AusAID management team
May 2013	Prepare review plan (Parts 1 and 2)	Review team
May 2013	Finalise review plan (Parts 1 and 2) (including endorsement by IEC)	AusAID management team reference group
Part 1: Quality review		
May –July 2013	Conduct quality analysis and prepare draft report (Part 1)	Review team
July 2013	Review draft report (Part 1) (including review by IEC)	AusAID management team reference group
Part 2: Synthesis		
July 2013	Revise review plan (Part 2)	Review team
July 2013	Finalise revised review plan (Part 2) (including endorsement by IEC)	AusAID management team reference group
July – September 2013	Conduct synthesis and prepare draft report (Part 2)	Review team
September – October 2013	Peer review of draft report (Part 2)	Reference group plus additional stakeholders/experts
October – November 2013	Prepare proposed final report (Parts 1 and 2)	Review team
Finalisation		
November 2013	Provide comments on proposed final report	IEC
December 2013	Prepare final version of final report	AusAID management team
December 2013	Publish report	AusAID management team

Annex 2: Detailed methodology

1 Overview of approach

Quality of Australian Aid Operational Evaluations is a meta-evaluation to assess the quality and credibility of the 87 independent evaluations completed in 2012 of aid initiatives implemented by the former AusAID.

The review sought to answer the following five questions:

1. What are the basic characteristics of different levels of independent evaluation in the aid program and the history and nature of independent evaluation at the initiative level?
2. To what degree do independent evaluations provide a credible source of evidence for the effectiveness of the Australian aid program?
3. What are the major strengths and weaknesses of independent evaluations of Australian aid initiatives?
4. What are the factors that contribute to their quality and utility?
5. What actions should be taken to improve the quality and utility of independent evaluations?

Table 1 provides an overview of the sources of evidence and the approach to analysis for each of these questions.

Table 1 Sources and approach to analysing evidence for the key review questions

Review question	Sources and approach to analysing evidence
1. What are the basic characteristics of different levels of independent evaluation in the aid program and the history and nature of independent evaluation at the initiative level?	Information on history of evaluation of Australian aid supplied by DFAT Interviews with key informants (selected DFAT staff and experienced consultants) Analysis of data on characteristics of evaluations and evaluated initiatives (extracted from AidWorks)
2. To what degree do independent evaluations provide a credible source of evidence for the effectiveness of the Australian aid program?	Quality review of evaluation reports Analysis of characteristics of evaluated initiatives to assess coverage of the Australian aid program
3. What are the major strengths and weaknesses of independent evaluations of Australian aid initiatives?	Quality review of evaluation reports Interviews with key informants (selected DFAT staff and experienced consultants)
4. What are the factors that contribute to their quality and utility?	Analysis of any 'statistical and qualitative associations' between evaluation quality and evaluation characteristics Interviews with key informants and selected DFAT evaluation managers & evaluators involved in a sample of evaluations
5. What actions should be taken to improve the quality and utility of independent evaluations?	Quality review of evaluation reports Interviews with key informants and selected DFAT evaluation managers and evaluators involved in a sample of evaluations Analysis of operational context (including teleconference with DFAT management team)

2 Detailed methodology

The review methodology comprised the following steps:

Step 1: Quality review of evaluation reports

The first step reviewed the quality of all initiative-level independent evaluations completed in 2012. The Review Team undertook a structured review of each of these 87 evaluations (including the associated terms of reference and the evaluation plan where available) using a set of 15 quality criteria.

The 15 quality criteria include nine general evaluation quality standards. These are based on a selection from the DFAT Aid Monitoring and Evaluation Standards, which draw on the internationally-agreed Organisation for Economic Cooperation and Development—Development Assistance Committee (OECD-DAC) evaluation quality standards. In developing these criteria we also drew on a number of existing frameworks for assessing the quality of evaluations.[23] These nine criteria were refined through a pilot test phase.

[23] L. Spencer, J Ritchie, J. Lewis, and L. Dillon (2003) *Quality in qualitative evaluation: a framework for assessing research evidence*, Government Chief Social Researcher's Office, Cabinet Office, UK; DFID (2012) Quality Assurance: Template for entry, Evaluation Department, UK ; DFID (2012) and Quality Assurance: Template for Exit, Evaluation Department, UK; UNICEF (2010) UNICEF-Adapted UNEG Quality Checklist for Evaluation Terms of Reference, July, Evaluation Office, UNICEF, New York; UNICEF (2013) Global Evaluation Reports Oversight System, January, Evaluation Office, UNICEF, New York.

The review team also assessed the quality of each evaluation's assessment against six standard Australian aid quality criteria: relevance, effectiveness, efficiency, impact, sustainability and gender equality.[24]

A standard pro forma (included at Annex 3) was used to record the assessments. Each criterion has specific guidance and definitions that the team applied in a systematic manner to each evaluation.

Using a six-point rating scale, each of the criteria was given a rating that captured our assessment of how well the evaluation addresses and provides evidence for each criterion. A short justification for each rating was given, while in the review documentation itself comments were added and highlighted text marked in order to record detailed evidence for our judgments. We used a coding system to organise and analyse the qualitative evidence. An assessment was also made as to whether numerical ratings, where provided, for the standard Australian aid quality criteria were robust.

The reviewer provided a short text summary for each quality area and identified whether the evaluation documents (evaluation report, terms of reference or evaluation plan) could be recommended as examples of good practice.

Step 2: Data analysis (stage 1)

Data analysis was conducted in two stages. Stage 1 involved conducting a preliminary analysis of data drawn from the quality review of the 87 evaluations. Stage 2 involved a deeper analysis of key enablers and inhibiters of evaluation quality.

Stage 1 in our data analysis brought together data from several sources: the numerical ratings from the quality review and the narrative data recorded to justify the scoring of each criteria; a set of explanatory variables for both the evaluation and the initiative that was evaluated; the ratings for the Australian aid quality criteria included in the evaluation reports; and previous initiative performance self-assessments (Quality at Implementation ratings).

Based on this data, the team examined patterns in the quality of the independent evaluations, as determined by our ratings for the 15 different quality criteria. The team did not present summary assessments based upon aggregate or average scores for an evaluation, but concentrated on analysing the range of scores for each criteria.

The ratings were analysed by two groups of 'explanatory' variables as set out in the terms of reference: (i) initiative characteristics (e.g. initiative value, country or region, primary sector) and (ii) evaluation characteristics (e.g. independent completion report / independent progress report, partner-led, cluster, evaluation team size, evaluation length). Several of these characteristics were extracted from the AidWorks database, along with others extracted during our review of the evaluation documents themselves. These possible explanatory variables were recorded in a spreadsheet, together with the review team's rating on the quality of the evaluation. Using this spreadsheet, the review team conducted statistical analysis of the data, in order to identify any correlations or patterns. These data were supported by narrative data recorded from the evaluation reports and coded to provide qualitative evidence and examples to illustrate and improve understanding of the quality criteria scoring.

---

[24] The robustness of the evaluators' numerical ratings against the standard M&E criterion was not considered in this review; however, the review does consider the evaluation's assessment of the adequacy and use of the initiative's M&E system under general evaluation quality standard criterion 5.

We also compared evaluation ratings against the Australian aid quality criteria (where provided) to the quality at implementation (QAI) ratings used by DFAT, to analyse differences and similarities between the two sets of ratings and identify any patterns and trends. However, given the limited data available there were no clear findings for this analysis, so it was not discussed in the final report.

Step 3: Interviews

The purpose of the interviews was to explore which factors have contributed to either the strengths or the weaknesses of evaluations (i.e. enabling or inhibiting factors) and, where appropriate, what actions can be taken to improve the quality and/or coverage of independent evaluations (key evaluation questions 4 and 5).

Qualitative evidence was collected during interviews with 10 Australian Government aid initiative managers and three senior executives, and 14 consultants who had responsibility for writing the evaluation reports. The selection of interviewees was based on the findings from our quality review of evaluation reports, and the preliminary findings of the analysis. We selected interview managers and evaluators for evaluations assessed as high or low quality across several criteria. This helped us identify common factors that appear to be either promoting or inhibiting successful evaluations. We also interviewed a small number of senior executive officers from program areas and the former PEPD, to obtain further evidence relating to key evaluation questions 1, 3, 4 and 5. While it proved marginally more difficult to secure interviews with Australian Government aid program staff than with consultants, we were in the end able to achieve a fair division in our interviews between the two groups.

Interviews followed as far as possible an 'appreciative inquiry' approach, where learning from positive experiences was emphasised. Questions were asked in a neutral and open-ended manner and respondents were encouraged to reflect on their experiences in a constructive, lesson-learning way.

Interviewees were assured that the records of all interviews would be confidential. Where requested by the interviewee, we provided general feedback on the strengths and weaknesses of their evaluation report. We did not enter into discussion on ratings for the different criteria.

An introductory email was sent to the interviewee in advance to explain the purpose of the interview and to give a list of possible questions that would be asked. Table 3 provides the indicative list of interview questions.

Table 3 Indicative interview questions

Questions for Australian Government aid initiative managers
1. What in your view were the particular strengths of the evaluation?
2. How useful/practical were the recommendations?
3. What worked well when you were managing the evaluation?
4. Were there any particular challenges in managing this evaluation?
5. If commissioning the same evaluation again, is there anything you might consider doing differently?
6. What actions could be taken within AusAID to improve the quality of evaluations?
Questions for evaluators
1. What were the key factors that assisted/facilitated your evaluation?
2. What were the major challenges you faced (if any) in undertaking this evaluation and how did you overcome them?
3. As an evaluator, are there any specific steps that AusAID could take to improve the quality of its evaluations?

Detailed interview notes were shared with the review team. The notes were recorded in a concise and consistent format categorised by question so that we could compare all the answers to a particular question. The analysis grouped responses around the main review questions, and also the particular interview questions aimed at the different types of respondent. These were then aggregated and coded around emerging types or categories of enabling or inhibiting factor (e.g. factors that are internal/external to DFAT or are related to particular aid modalities or forms of evaluation). The coded data was recorded in a spreadsheet to facilitate analysis.

For validation purposes, the review team held an internal half-day meeting to discuss the evidence and the results of the analysis to agree on the key findings.

Step 4: Analysis (stage 2)

The second stage in our analysis focused on key evaluation questions 3 to 5 but particularly on gaining a deeper understanding of the underlying drivers and inhibiters of quality. We triangulated between the different forms of evidence, including the analysis from stage 1, qualitative evidence from the evaluation reports and the interviews, and other contextual information. We looked for consistent themes across the evidence, and developed hypotheses about why certain aspects of evaluation are undertaken well and others less so, and what factors positively or negatively affect evaluation quality.

A workshop took place by teleconference at which we presented our initial analysis and findings to the DFAT management team for the *Review of Australian Aid Operational Evaluations*. At this meeting we identified and discussed the patterns and themes emerging from our analysis, identified any gaps in our data and discussed how these could be filled, and discussed possible hypotheses for further investigation. We discussed the strength and salience of the key themes and findings (particularly with respect to the Australian Government's operational context for aid), and any further analysis required.

Step 5: Identifying good practice

During the review, examples of good or model practice were identified to support ongoing learning and improvement within DFAT. We selected examples that cover the evaluation report, terms of reference and evaluation plan.

3 Quality assurance

To ensure consistency in the quality review, a number of measures were taken. The initial design of the quality assessment was tested during the planning stage within the team on a sample of three evaluations to improve the format and strengthen mutual understanding within the team.

The team leader oversaw the quality assurance process to ensure consistent review standards. After the first week, an initial quality assurance review took place where a sample of the initial set of reviews was re-assessed by different reviewers within the team. Thereafter, a sample of reviews was quality assured each week. Over a four-week period, the team leader cross-checked 20 evaluations, and the other two team members quality assured a total of 12, making a total of 32 or 37 per cent of the total number of evaluations reviewed.

The team leader also regularly checked the consistency of the data entered into the spreadsheet by the team members.

4 Limitations

The key limitations to the methodology were:

- The limited availability of evaluation plans and, to a lesser extent, terms of reference constrained the evidence base for the quality review. Because of this, the review team focused primarily on the quality of the evaluation reports and judged their quality irrespective of the availability of the terms of reference or evaluation plan. However, where terms of reference and evaluation plans were available, they were reviewed to identify examples of good practice.
- Each evaluation was assessed against a set of generic criteria without taking into account the intended purpose or utility of the individual evaluation. Evaluations can be undertaken for a range of purposes, as reflected in the variation in scope, design and resourcing. However, it was not within the scope of this review to assess whether the evaluations were fit for the purpose intended, and we thus reviewed the quality of the evaluations against this set of nine generic evaluation quality standards. It is reasonable to expect that all evaluations would at least minimally meet these criteria (with the possible exception of 'assessment of intervention logic'). This also includes joint or partner-led evaluations.

5 Conflict of interest

A potential conflict of interest may have arisen if the team were to review an evaluation or an initiative that either ITAD Ltd or individual review team members were involved in designing, implementing or evaluating. However, this did not occur.

6 Ethical conduct

ODE's evaluations are guided by relevant professional standards, including the Australasian Evaluation Society's guidelines for the ethical conduct of evaluations.

ITAD adopts accepted standards and ethical principles for the conduct of evaluations. ITAD is a corporate member of the UK Evaluation Society (UKES) and the International Development Evaluation Association (IDEAS), and adopts the UKES *Guidelines for good practice in evaluation* and the IDEAS *Competencies for development evaluation evaluators, managers, and commissioners*. ITAD recognises the United Nations Evaluation Group's *Ethical guidelines for evaluation,* the UK Department for International Development's *Ethics principles for research and evaluation* and the OECD Development Assistance Committee's *Quality standards for development evaluation*. For this review, ITAD observed the Code of Ethics of the Australasian Evaluation Society.

Annex 3: Pro forma for quality review

This pro forma was used by the ITAD review team to assess and rate the 87 evaluations against the 15 quality criteria.

Cover sheet

ITAD reviewer: [ ]

Initiative name: [ ] Country: [ ] Sector: [ ]

Date of evaluation: [/ / 2012]

Type of evaluation: [ ] Partner-led (joint): y/n Cluster: y/n

Evaluation team leader: [ ] Evaluation team composition:[ ]

Total days allocated: [ ] Fieldwork days: [ ] Total person-days: [ ]

Ratings

Satisfactory		Less than satisfactory	
6	Very high quality	3	Less than adequate quality
5	Good quality	2	Poor quality
4	Adequate quality	1	Very poor quality

Not covered: The criterion was not included in the evaluation

Not assessable: The criterion was included but it is not possible to assess quality because there is too little information

Key quality areas and criteria	Quality statements	Rating 1–6	Evidence
Evaluation purpose and scope			
1 Purpose of evaluation	The purpose of the evaluation is provided, including the overall purpose and primary users of the information		
2 Scope of evaluation	The scope matches the evaluation resources; methods are defined and roles of the team, AusAID management and others are set out.		
Overall comments			
Evaluation methodology			
3 Assessment of Intervention logic	The evaluation assesses the intervention logic or theory, or equivalent, including underlying assumptions and factors. The report assesses the clarity of initiative objectives		
4 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) Limitations to the methodology and any constraints encountered are described Ethical issues such as privacy, anonymity and cultural appropriateness are described and addressed		
5 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.		
Overall comments			
Findings, conclusions and recommendations			
6 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.		
7 Evaluation questions and criteria	The report identifies appropriate evaluation questions and then answers them. Any ratings, if given, are justified. Where this is not done, explanations are provided. 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. 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		

9 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		
Overall comments			

Application of standard Australian aid quality criteria				
	Was this criteria assessed in the report? y/n	If, yes, what was the rating given in the evaluation? 1–6 or no rating or score given (N/S)	Our own rating of the quality of the assessment (Does the report provide evidence of the adequate application of the criteria?) 1–6 or N/C: (not covered), N/A (not assessable)	Is the rating given in the evaluation robust? (rated too high (+), too low (–) or robust (N/R) or not assessable N/A)
1 Relevance The initiative is the most appropriate way to meet high priority goals that Australia shares with its development partners within the given context.			Evidence: Rating: [ ]	Comments: Assessment:
2 Effectiveness The report provides evidence of the adequate application of the criteria of 'effectiveness'. The initiative is meeting or will meet its objectives, and is managing risk well.			Evidence: Rating: [ ]	
3 Efficiency The resources allocated by Australia and its partners are appropriate to the objectives and context, and are achieving the intended outputs. Value for money or cost-effectiveness looks beyond how inputs were converted into outputs, to whether different outputs could have been produced that would have had a greater impact in achieving the project purpose.			Evidence: Rating: [ ]	
4 Impact Impact looks at the wider effects of the project—social, economic, technical, environmental—on individuals, gender, age groups, communities, and institutions.			Evidence: Rating: [ ]	
5 Sustainability Significant benefits will endure after Australia's contribution has ceased, with due account given to partner systems, stakeholder ownership and plans for phase out.			Evidence: Rating: [ ]	
6 Gender equality The initiative incorporates appropriate and effective strategies to advance gender equality and promote the empowerment of women and girls.			Evidence: Rating: [ ]	

Good practice example?		(y/n)	Evidence
	Evaluation report Does the evaluation report represent a good example of evaluation practice, and if so why and in what areas?		
	Evaluation plan If the evaluation plan is available, does it provide a good example of a detailed plan to conduct the evaluation?		
	Terms of reference If the terms of reference are available, do they provide a clear background and rationale, and a list of prioritised evaluation questions?		
Follow-up interview recommended?	(y/n)	Reason:	
	Topics/questions to be asked if this evaluation is chosen for interview		

Annex 4: Additional analysis

Disaggregated results of quality review

Figure 8 Disaggregated results of our assessment of 87 evaluations against 15 quality criteria

Quality review rating	Very poor (1)	Poor (2)	Less than adequate (3)	Adequate (4)	Good (5)	Very high (6)	Not assessable[25]	Total no. assessed	% assessed as adequate or better (4–6)
Evaluation quality standards									
1. Purpose for evaluation	1	2	8	42	27	7	0	87	87%
2. Scope of evaluation	2	9	21	29	21	4	1	86	62%
3. Assessment of intervention logic	11	17	20	19	15	1	4	83	42%
4. Appropriateness of methodology	6	21	25	11	19	5	0	87	40%
5. Adequacy and use of M&E	2	5	11	25	34	9	1	86	79%
6. Context of the initiative	1	8	24	21	27	6	0	87	62%
7. Evaluation questions and criteria	2	7	13	35	26	4	0	87	75%
8. Credibility of evidence and analysis	2	5	16	33	26	5	0	87	74%
9. Recommendations	0	6	8	24	41	8	0	87	84%
Application of standard Australian aid quality criteria									
1. Relevance	0	0	16	24	26	5	16	71	77%
2. Effectiveness	0	6	18	20	31	7	5	82	71%
3. Efficiency	4	12	20	20	15	7	9	78	54%
4. Sustainability	4	6	23	15	17	9	13	74	55%
5. Impact	7	7	10	9	11	1	42	45	47%
6. Gender equality	2	13	12	12	19	5	24	63	57%

[25] 'Not assessable' was recorded where there was insufficient information in the evaluation report to make a judgment. Where the report did not address a particular criterion because it was not (or did not appear to be) included in the scope of the evaluation, no assessment was made. There were two other cases where a criterion could not be assessed for other reasons: one evaluation (INI691) covered 13 different trust funds and assessing intervention logic for all was unfeasible; the terms of reference for the other evaluation (INI426) were not available and the scope of the evaluation could not be assessed based on the information in the evaluation report alone.

Degree of association between the 15 quality criteria

Figure 9 presents analysis showing the degree of association between the 15 quality criteria. The higher the correlation coefficient, the stronger the statistical association. For example, a score of 1 would indicate a perfect correlation (all scores exactly the same). A score of –1 would indicate a perfect negative correlation. Correlations of 0.4 or higher are shaded.

This analysis indicates that 'credibility of evidence and analysis' is the criterion that is most strongly associated with the other criteria. It is therefore the best predictor of quality—if an evaluation received a good rating for this criterion, then it is also likely to have good ratings for many of the other criteria. This is why we have frequently used the findings against this criterion as a proxy for overall evaluation quality in our analysis.

Figure 9 Correlation analysis showing the degree of association between criteria

	1.Purpose	2. Scope	3. Assessment of intervention logic	4. Methodology	5. M&E	6. Context	7. Evaluation questions & criteria	8. Credibility of evidence & analysis	9. Recommendations	1. Relevance	2. Effectiveness	3. Efficiency	4. Impact	5. Sustainability	6. Gender
Evaluation quality standards															
1. Purpose	1														
2. Scope	0.17	1													
3. Assessment of intervention logic	0.36	0.27	1												
4. Methodology	0.43	0.41	0.33	1											
5. M&E	0.16	0.25	0.27	0.33	1										
6. Context	0.28	0.23	0.38	0.33	0.24	1									
7. Evaluation questions & criteria	0.31	0.18	0.25	0.35	0.33	0.36	1								
8. Credibility of evidence & analysis	0.40	0.30	0.37	0.47	0.39	0.39	0.58	1							
9. Recommendations	0.31	0.30	0.25	0.34	0.31	0.42	0.40	0.63	1						
Application of standard Australian aid quality criteria															
1. Relevance	0.38	0.29	0.27	0.33	0.26	0.41	0.20	0.46	0.28	1					
2. Effectiveness	0.23	0.12	0.17	0.31	0.35	0.30	0.49	0.67	0.45	0.36	1				
3. Efficiency	0.28	0.32	0.35	0.29	0.32	0.36	0.35	0.60	0.48	0.30	0.43	1			
4. Impact	0.10	0.21	0.29	0.36	0.37	0.33	0.30	0.55	0.41	0.34	0.28	0.44	1		
5. Sustainability	0.16	0.27	0.33	0.25	0.24	0.09	0.38	0.66	0.47	0.34	0.56	0.41	0.50	1	
6. Gender equality	0.24	0.21	0.30	0.37	0.48	0.22	0.17	0.48	0.31	0.37	0.37	0.27	0.54	0.35	1

Enabling and inhibiting factors affecting evaluation quality

Figure 10 Partner-led evaluations and evaluations managed by the former AusAID alone—quality comparison

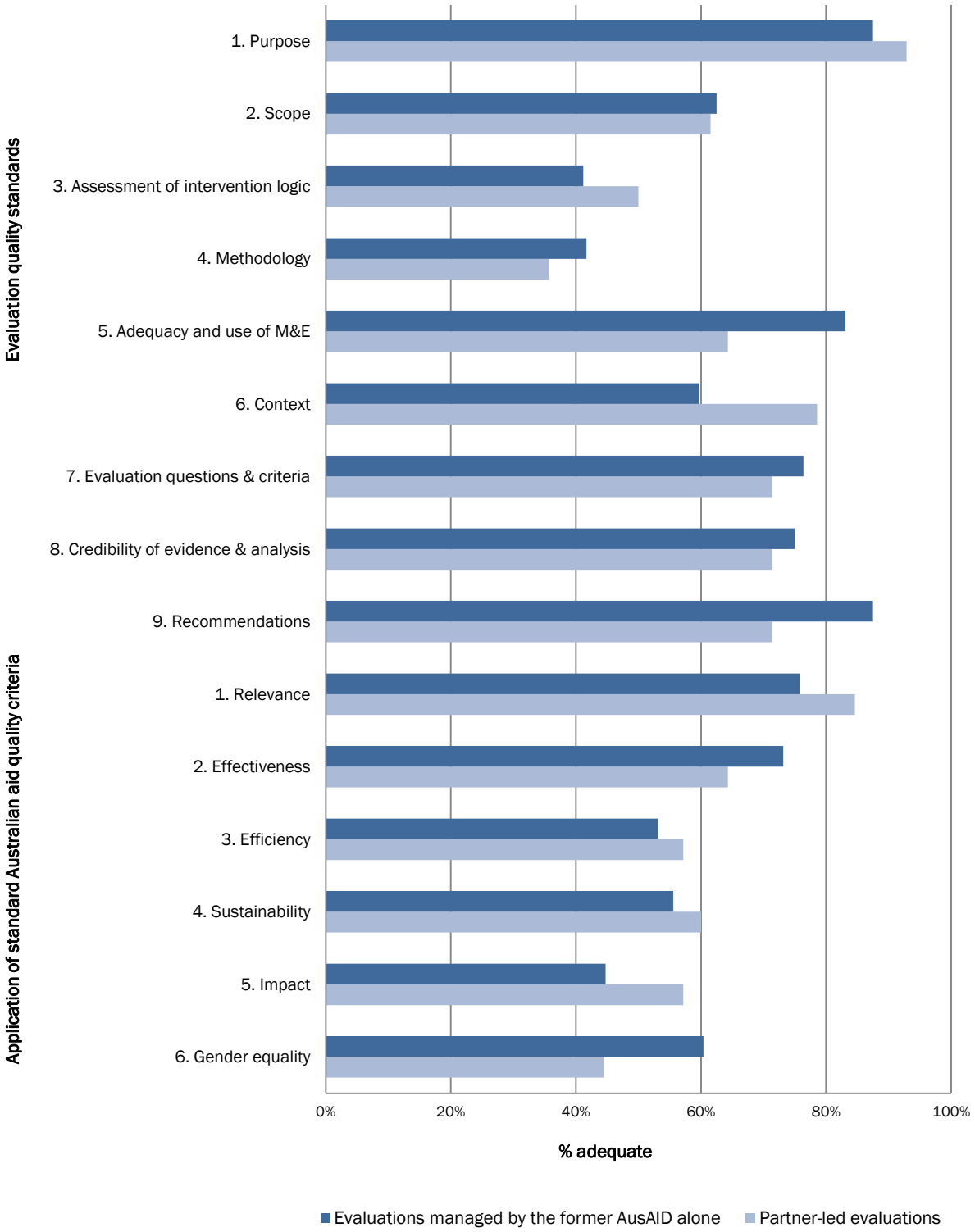


Figure 11 Credibility of evidence and analysis by primary sector

Sector	Assessment for credibility of evidence and analysis							
	Total	Very poor	Poor	Less than adequate	Adequate	Good	Very high	% adequate or better
Overall[26]	**87**	**2**	**5**	**16**	**33**	**26**	**5**	**74%**
Human rights	10			5	3	2		50%
Food security and rural development	8	1		2	2	3		63%
Water and sanitation	6	1		1		3	1	67%
Security and justice	6			2	2	1	1	67%
Improved government	15		1	3	7	3	1	73%
Health	8		2		2	4		75%
Education	10		1		6	3		90%
Infrastructure	5				3	2		100%
General development support	7				4	3		100%
Humanitarian response	5				3	2		100%

[26] Sectors with fewer than five evaluations have been omitted from this table, to discourage comparisons based on a very small sub-sample. These sectors are Business, finance and trade, Environment and natural resource management, and Conflict prevention and resolution.

Annex 5: DFAT Aid Monitoring and Evaluation Standards

The lessons for evaluation commissioning areas presented in this review should be read in conjunction with the DFAT Aid Monitoring and Evaluation (M&E) Standards,[27] which articulate expectations of the quality expected from a range of M&E products.

The Standards provide a useful resource for evaluation commissioning areas. The Standards can assist DFAT officers to articulate consistently their requirements to M&E practitioners and the industry more broadly; to assess and assure the quality of the M&E products they receive; and to work with implementation teams and M&E practitioners to improve the quality of products where necessary. Equally, the suppliers of M&E products benefit from this clear articulation of what is required, and the Standards provide a strong basis for the negotiation of the delivery and resourcing of quality products.

Developed by the former AusAID, the Standards had not been formally adopted at the time the evaluations in this review were commissioned and undertaken, but they were integrated into agency evaluation guidance in 2012.

The Standards in this series are:

Standard 1: Investment Design (required features for M&E)

Standard 2: Initiative M&E Systems

Standard 3: Initiative Progress Reporting

Standard 4: Terms of Reference for Independent Evaluations

Standard 5: Independent Evaluation Plan (Methodology)

Standard 6: Independent Evaluation Report

Standard 7: Monitoring Visits

Standards 4, 5 and 6 relating to independent evaluations are presented here.

[27] The full DFAT Aid Monitoring and Evaluation Standards can be accessed at: http://aid.dfat.gov.au/Publications/Pages/monitoring-evaluation-standards.aspx

Standard 4: Terms of reference for independent evaluations

Note: The term evaluation is used in this document to refer to both reviews and evaluations. Evaluation would normally refer to a piece of work with a higher degree of methodological rigour usually requiring longer time frames and additional resources.

Background and orientation to the evaluation

4.1. A brief orientation to the initiative is provided

As the terms of reference are used to explore with proposed consultants whether or not they are interested in, or to comment on the proposed evaluation, the orientation must ensure the TORs are a stand-alone document. Important information includes: the total value; the time frame; a summary of the expected end-of-program outcomes; a short summary of the key approaches employed (such as training, technical advisers, secondments, provision of infrastructure, equipment, and budget support or pooled funding etc.). The context in which the initiative is situated is described such as the program strategy and/or delivery strategy that the initiative aims to address as well as the partner government development plans of relevance. The delivery mechanism is described (contracted, multi-lateral development partner, NGO) and whether or not the initiative is a project, program or facility. Any information which can guide the reader in quickly understanding the scope/reach of the initiative is provided.

4.2. The purpose of the evaluation is described

The TOR clearly identifies the overall purpose(s) and shows which purposes are of most importance—accountability, initiative improvement, knowledge generation, or developmental.[28] This allows the consultant to reflect these priorities in the evaluation plan. The primary users of the information are identified so that the consultant can collect relevant information, contribute to deepening an understanding of the findings during the mission, and prepare an appropriate report. Primary users 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.

There is a wider audience for evaluations than the primary users, and the reports are usually published on the DFAT website.

4.3. The TOR identifies the key decisions (management, operational and/or policy) which the evaluation is intended to inform

Any important management decisions that the primary users are expected to make are identified and described. Management decisions are more specific than the purpose and involve decisions such as whether or not to extend an initiative, whether or not to involve a new partner, whether partner systems are ready for use, or whether to consider a new modality for a future initiative.

[28] Developmental evaluation is used in highly complex situations, or in programs that are in the early stages of innovation. See Gamble (2008) A Developmental Evaluation Primer. JW McConnell Family Foundation.

4.4. Key Issues are identified and discussed

Any important issues that have informed the call for, or design, of the evaluation terms of reference are identified and described. They are described in neutral language and do not infer an expectation of findings. They are described in sufficient detail to enable the evaluator to develop the evaluation plan to adequately explore the issues.

Key evaluation questions or scope

4.5. The key evaluation questions are consistent with the overall purpose and management decisions of the evaluation

Each of the key evaluation questions is clearly related to the stated purpose(s) of the evaluation (and clearly related to the key management decisions). There are an adequate range of questions to meet all the stated purposes, and to ensure DFAT's information needs are met. There are no additional questions unrelated to the stated purpose. Although the DAC criteria are an important consideration for the evaluation, these have not been cut and pasted into the TOR resulting in broad questions of ambiguous scope. There is a single list of questions in one place in the TOR.

4.6. Priority evaluation questions are identified

Of the full list of questions, the TORs clearly show what the priority questions for the evaluation are. This will allow the evaluator to make judgments during the evaluation of what questions must be answered in the final report, and what questions would be desirable. These priorities are consistent with the overall purpose of the evaluation. Ideally, only priority questions are posed, but in some cases where stakeholders have generated numerous questions which they want to keep in the TORs, prioritising these can be a way of showing the evaluator exactly what are the critical questions.

4.7. The scope of the questions is suitable for the time and resources available for the evaluation

Typically, a 12-day in-country evaluation can only address four or five broad questions. During the development of the evaluation plan, these are broken down into a larger number of sub-questions or information requirements. In addition, for a typical interview with a respondent without the need for translation, only a small number of topics can be addressed with depth. Still, this is only possible if both the interviewer is skilled in questioning techniques and the respondents are relatively articulate and experienced in the topic areas. In addition to collecting the information, it also needs to be processed, interpreted and reported on. The questions posed in the TORs reflect this reality.

4.8. Sufficient supporting information is provided about Key Evaluation Questions to guide the development of an appropriate evaluation plan

Evaluation questions are not broad or vague or open to a wide range of interpretations. There is clarity in either the Key Evaluation Questions, or the supportive information provided. The evaluator will be able to break down questions and identify the specific information requirements. For this to be successful and for the purpose(s) to be met, the evaluator will need to be able to correctly interpret the expected information from the way the questions are worded. The Key Evaluation Questions (and supportive information) pose questions in a way that the evaluator can select suitable methods for the time and resources available (for example, cause-and-effect questions are difficult to answer in a short review without access to suitable secondary data sources).

Evaluation process

Adequate time and resources are required to enable the evaluation to be completed with an adequate degree of rigour. The following processes are allowed for:

4.9. A verbal briefing of the key issues and priority information is planned

A phone or face-to-face briefing is planned to discuss the background, issues and priorities for the evaluation with the evaluator before the evaluation plan is developed. Sufficient time must be allocated to allow DFAT and the evaluator to work together to clarify scope, priority questions and issues, and general approach to methods. This may require more than one discussion.

4.10. Adequate time has been allocated for document review and document appraisal

Time has been allocated to reviewing initiative documentation (approx. 2 days) as well as time to appraise any key documents such as strategies or the M&E system (often a day per document for full appraisal).

4.11. There is a requirement for an elaborated evaluation plan—the depth of planning required reflecting the importance of the review/evaluation questions and management decisions

The depth of planning required for an evaluation reflects the importance of the related decisions that will be made in response to the evaluation. If important decisions are to be made then more time is allocated. Typically for a DFAT-commissioned evaluation three days is required to develop an evaluation plan which includes a fully elaborated methodology. See Standard 5: Evaluation plan for more details.

4.12. The submission date for the evaluation plan allows sufficient time for data collection activities to be scheduled

The data collection activities proposed by the evaluation team will be set out in the evaluation plan. This plan needs to be submitted to the evaluation manager well in advance of the in-country visit to allow for data collection activities such as interviews and site visits to be scheduled.

4.13. Proposed scheduling allows for adequate data collection, processing and analysis to answer Key Evaluation Questions

The proposed schedule in the TOR is not too detailed as this is developed after the evaluation plan identifies suitable respondents and activities to address the evaluation questions. There are a 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 and identify further requirements as the mission unfolds.

4.14. A feedback session to relevant information users are planned together or separately depending on the sensitivity of findings (e.g. Aide Memoire, discussion or presentation)

There is adequate time to provide detail in evaluation findings to allow contestability of those findings, and feasibility of recommendations with key stakeholders. As the uses of information may be different for the different primary users, a suitable range of feedback options are offered.

4.15. There is provision for processing the information collected to enable systematic analysis and interpretation, and the development of an evidence base

The evaluation team has been given adequate time to process information from interviews, document reviews and appraisals, observations or other methods to provide a credible evidence base to support findings. Typically, three days would be required for processing of data for a 12 day in-country mission that relied strongly on interviews. More complicated evaluations (or those with an emergent design) would require more time. This is additional time to actual report writing. Flexibility is balanced with value for money, but final time frames should be negotiated with the evaluator.

4.16. Adequate time is made available to complete the draft report

The number of days allocated to completing the 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 on the final draft. It is recommended to allow the evaluation team sufficient time to rest after the mission and to reflect on the mission. For example, ten days allocated to report writing could require a three week period to deliver. It is also useful to discuss with evaluators whether or not they expect to be working on other reports and missions during this time.

4.17. The process for commenting is efficient and allows independence of the evaluation team final report

The process for commenting on the draft report is described and is efficient. Only relevant individuals are invited to comment, and the focus of their comments is identified. Note that those invited to comment on the final report would also be invited to comment on the evaluation plan to ensure their final comments are within the scope and expectations for the evaluation. The TOR explains that DFAT will either provide comments in a consolidated form to the evaluation team, or, allows additional time to respond to a large number of comments from all stakeholders.

Note: Be aware that where DFAT personnel consolidate comments, there must be transparency of decisions on what comments to include or remove. It may be necessary to provide comments from different stakeholders (national partner, implementation team and DFAT) separately if there are conflicting views.

4.18. Adequate time has been allocated to responding to comments

The time allocated to the evaluation team to respond to comments reflects a) the likely range of comments generated; and b) the possibility that comments require significant structural change in the final report.

4.19. The roles and functions of each team member are stated

Although it is the responsibility of the team leader to produce the final report and provide detailed direction on tasks in the evaluation plan, the TORs show DFAT expectations about how each team member will contribute. This is especially important with respect to writing responsibilities. There is clear guidance on the extent to which DFAT expects international and national consultants to participate and assume responsibility for particular tasks. If there is any requirement of the team leader for capacity building of team members, then adequate time has been allocated to carry this out effectively.

4.20. Skill sets of evaluation team reflect priority questions

Unless there is a compelling reason provided, the team leader is an evaluation expert, not only technical expert in the relevant sector or thematic area. They are supported by technical specialist(s) who will focus on technical aspects of the evaluation. The development of the evaluation plan is allocated to the team leader who will be responsible for its implementation. The development of the evaluation plan is not allocated to a team member. The tasks and balance of work of technical advisers reflects the evaluation questions. For example, if there is a strong focus on gender, or initiative management systems, then the number of days allocated to technical specialists from other sectoral areas reflects this focus.

4.21. The reporting requirements allow DFAT to track progress of the evaluation without distracting the team from carrying out important evaluation activities

The requirement for reports during the mission provide for a good balance between monitoring the progress of the evaluation with allowing the team to focus on important evaluation activities. The evaluation plan is a critical document for DFAT to ensure that the evaluator has correctly interpreted the TORs and has made suitable plans to conduct the evaluation to meet the TORs and reasonable standards of rigour. Other reporting requirements to consider are a) the aide memoire (which should be short and only provide anticipated key findings and recommendations); and b) the provision of any processed data. If requiring the provision of processed data, it is important to consider the implications for preserving the confidentiality of respondents.

Note: A negotiation of the TORs should be encouraged during the contract negotiations. This allows the team leader to provide professional advice on the feasibility of the TOR in terms of the scope of questions and the resources applied.

Standard 5: Independent Evaluation Plans

Note: The Evaluation Plan is developed by the evaluator based on the ToR. It is a negotiated document between the client and the evaluator and should provide more detail and reflect final agreements after that negotiation. The evaluation plan should be submitted as early as possible, to enable scheduling of site visits, interviews and other data collection activities. The agreed Evaluation Plan ought to provide the basis by which evaluator performance is assessed.

5.1. The evaluation plan is based on a collaborative approach

The evaluator has consulted DFAT, and the stakeholders identified as important by DFAT, to develop the evaluation plan. Consultation may have been in-person, by phone or by email. Important stakeholders have been given the opportunity to comment on the evaluation plan before the evaluation commences. Note: This ensures that additional information will not be requested after the data collection phase is complete.

5.2. Primary intended users of the evaluation are clearly identified

An evaluation cannot meet the needs of all interested stakeholders. Individuals (by title) in named organizations should be identified as the primary users of the evaluation findings. These are the people who will be using the information to make judgments and decisions. Audience is a different concept and often refers to a broader group of people that may be interested in, or may be affected by any decisions that result from the evaluation.

5.3. The purpose and/or objectives of the evaluation are stated

These would normally be taken from the terms of reference. The evaluation design restates these so that the evaluation plan is a stand-alone document.

5.4. A summary is provided to orient the reader

This is an introductory orientation of the overall design of the evaluation. It is short, about one paragraph in length. For example, it could highlight whether the evaluation is predominantly exploratory or descriptive, or whether a cause and effect design is proposed, or whether or not any case studies would feature in the overall design. It would highlight the major methods for data collection and analysis. This is called the investigatory framework in research and evaluation terms. The evaluation plan does not go straight into detailed descriptions of methodology without this general orientation.

5.5. Limitations or constraints for the evaluation are described

The time available for the evaluation has implications for the scope of the evaluation. If a large number of questions are posed, but DFAT only wants a cursory look at many of these, then a shorter time frame may be appropriate. The evaluator highlights any important limitations in terms of time available, resources applied, or the expertise of the evaluation team to deliver a credible, defensible evaluation product. Political sensitivities are highlighted where appropriate. The implications of these limitations are discussed.

Note: A long list of limitations is not considered a substitute for a poorly negotiated TOR.

5.6. The Key Evaluation Questions are supplemented by detailed descriptions and/or sub-questions

Although the terms of reference is where DFAT communicates what the evaluation is to address, the evaluator will still need to give careful consideration to how these larger questions will be addressed. This means that more detailed information requirements and/or sub-questions are generated. Commonly, questions presented in a terms of reference are broad, therefore this more detailed information allows information users to know how the evaluator has interpreted the broader questions, and whether or not the evaluation will generate sufficient information to meet these broader questions. It also allows the DFAT evaluation manager to see the implications of the scope of the evaluation described in the terms of reference. This breakdown of information requirements or questions allows the reader to assess whether or not the original scope was realistic. The evaluation manager needs to pay careful attention to this aspect of the evaluation plan.

5.7. High-priority questions are identified

DFAT evaluations often have a very large number of evaluation questions that cover a very wide number of aspects of the initiative to be evaluated. Some of these questions will be more important than others. The evaluation plan reflects where the emphasis will be placed, and it is clear that DFAT's information needs will be met. The evaluation team will not usually be able to answer all the questions listed for all respondents and so will need to make decisions during interviews about what will be dropped and what is essential. The evaluation manager needs to be confident that the evaluator will, at a minimum, deliver information on the priority questions.

5.8. There is sufficient flexibility to be able to address important unexpected issues as they emerge

This flexibility may be built in to the questioning technique employed during an interview. It may be built into the schedule as a whole to allow new issues to emerge and be responded to through additional data collection if they are important. Where new issues cannot be adequately addressed within the schedule, there are processes to review possible trade-offs to allow them to be addressed.

5.9. Methods for each evaluation questions are described

The evaluation plan shows how each of the evaluation questions will be answered by describing the methods that will be used to collect the information. For most DFAT evaluations this is likely to include in-depth interviews, focus group discussions/interviews, document reviews and in some cases observations of activities. Large workshops are not usually a suitable method to gather substantive, reliable and valid information—however, they may have other important political purposes.

For several questions there may be a number of data collection methods proposed to strengthen confidence in the findings.

The design of major evaluation activities/studies should be annexed and include tools such as interview guides or questionnaires. In some cases the evaluator will need to develop these later, or adjust them as the evaluation proceeds, but there is an absolute expectation that the evaluator uses tools to guide each evaluation activity, and do not rely on memory of all the evaluation questions identified in the evaluation plan. Where team members are working in different locations then tools are available ahead of those evaluation activities so that data is collected systematically. If flexibility on this is required, then a compelling rationale is provided. Summary statements of methods that are not linked with specific evaluation questions are not considered adequate.

5.10. Methods are appropriate for the evaluation questions posed

Although this takes evaluation expertise, it is still worth reviewing the questions posed and considering if the methods described could reasonably answer the questions. For example, a focus group discussion would be most unlikely to answer a sensitive question; a review of a program strategy document (such as gender) would be unlikely to tell you if the initiative's actual gender activities were of a high quality. It would need to be supported by information from other sources.

5.11. Triangulation of methods is proposed

Triangulation is the use of a range of methods and/or sources of information to come to a conclusion or result. It can develop greater confidence in a finding. Given the short timeframe of most DFAT evaluations or reviews, it is difficult to employ a wide range of methods. To deal with this, the evaluation has planned to discuss similar questions across a range of different respondents within and across different organisations, 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.

5.12. Sampling strategy is clear and appropriate

Most evaluations will require some sort of sampling strategy across individuals, sites or time periods. 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. Where statistical inference

will be used to generalize from the sample, random sampling strategies are appropriate—especially stratified random sampling which reduces the sample size required.

5.13. The plan describes how data will be processed and analysed

The evaluation plan describes how the data will be processed, including measures to check and correct any errors in data, ensure security of storage and prepare for analysis. The plan also describes how the data will be analysed in order to answer the Key Evaluation Questions. This may not necessarily require advanced analytical methods, but users of the information can determine exactly what is to be done.

5.14. The plan identifies ethical issues and how they will be addressed

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, including when providing a database of the evidentiary basis to DFAT as part of the deliverables. Other relevant ethical issues are addressed including processes for reporting serious issues if identified during data collection.

5.15. The process for making judgments is clear

The evaluation plan makes it clear that DFAT requires the evaluator to make a professional judgement based on the evidence gathered and the agreed basis by which judgements are made (such as criteria or standards). DFAT's response to the evaluator's judgement should be provided in the Management Response to an evaluation. In some exceptional cases, DFAT may require an evaluator to report neutrally on facts and leave DFAT to make the final judgements, in which case the plan should make it clear how evaluative judgements will be made and by whom, as this is an important distinction and can affect the way information is collected and presented.

5.16. Approaches to enhance utilisation of findings are outlined

The importance of utilisation of findings needs to be communicated to the evaluator. There are a variety of well-tested approaches to utilisation that a professional evaluator will be familiar with (e.g. stakeholder engagement strategies for evaluation design or developing acceptance of recommendations before the report is published). Approaches to utilisation of findings are outlined in the evaluation plan. Utilisation begins with the evaluation design stage.

5.17. Scheduling guidance is provided

The schedule is developed by DFAT after the evaluation plan is submitted, and reflects guidance from the evaluator. The most common problem is that the persons recruited for interview are not always the best respondents for the evaluation questions posed. Often there are many donor meetings where respondents cannot provide substantive comment on many of the evaluation questions. Also consider the time for each interview with the associated evaluation questions. Most 60-minute interviews with a respondent cover no more than four or five key topics; less if translation is required. Sufficient time is available to meet with the implementation team. As part of reviewing the methodology DFAT negotiates the proposed list of respondents before final scheduling. The evaluator scheduling guidance is realistic for the timeframe. Sufficient time is allocated to other methods proposed. There is sufficient time allocated to evaluation team discussions and early data processing (not late at night).

5.18. Evaluation tasks are allocated to team members

It is very important that each team member knows before the evaluation begins what they will be expected to do. It is not appropriate for the team leader to allocate reporting responsibilities on the last day of the in-country mission. The evaluation plan shows what responsibilities each team member has so they can ensure that adequate data is collected, processed, and interpreted and they can meet a high standard during the reporting stage. It is often useful to show which evaluation questions each team member carries responsibility for.

Standard 6: Independent Evaluation Reports

Introductions

6.1. The background provides adequate information for individuals not familiar with the initiative

The background provides adequate information to enable individuals not fully familiar with the initiative to interpret the report. It summarises: the total value of the initiative; the number of years of the initiative; the stage of initiative implementation; the delivery mechanism; key expected outcomes of the initiative; and the key issues identified in the terms of reference.

6.2. A brief summary of the methodology employed is provided

Although a fully elaborated methodology was developed before the evaluation, a summary of the significant details is included. Sufficient information is required to enable the reader to quickly understand the evidentiary basis of the evaluation. The evidentiary base must be convincing and in proportion to the resources invested in the evaluation. The full methodology is annexed. Important aspects of the strategy to ensure findings are utilised are summarised here.

6.3. Key limitations of the methodology are described and any relevant guidance provided to enable appropriate interpretation of the findings

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.

6.4. The executive summary 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. 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).

Findings and analyses

6.5. The evaluation report clearly addresses all questions in the Terms of Reference / Evaluation Plan

Note: As the Evaluation Plan supersedes the Evaluation Terms of Reference, the Plan is the appropriate document to assess whether or not the evaluation has delivered on expectations. In the absence of an Evaluation Plan, the Terms of Reference should be used.

It is relatively easy to identify where each of the questions in the Evaluation Plan are addressed. The report does not need to be a mechanical presentation of these questions, but it should be relatively easy to negotiate the report and find relevant information about specific questions in the Evaluation Plan. 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.

6.6. The relative importance of the issues communicated is clear to the reader

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 breadth of description, depth of analysis and attention in the recommendations can indicate the degree of priority. The author may simply state the relative importance of issues.

6.7. There is a good balance between operational and strategic issues

The report addresses the full range of issues identified in response to the TOR and other critical issues that have emerged. There will be technical, managerial or operational issues that are very important to consider and are often at the core of many important challenges. The strategic direction or any higher order issues of the initiative have been given adequate space, and minor technical issues are treated in a more limited fashion. Flexibility is required where the TOR evaluation questions demonstrate that this balance was not required.

6.8. The report clearly explains the extent to which the evidence supports the conclusions and judgments made

For key findings, the basis of the findings and related conclusions is communicated clearly. This includes reporting the degree to which views are shared across respondents. The information is brought together from a range of sources, but communicated as a coherent whole. Evaluator opinions that are based on limited evidence are proposed as suggestive only.

6.9. Alternative points of view are presented and considered where appropriate

Alternative views must be 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.

6.10. Complicated and complex aspects of issues are adequately explored and not oversimplified

The report adequately acknowledges complicated aspects of issues, such as multiple contributing factors, or emergent challenges and opportunities. The report does not present simple solutions to these types of situations. The findings are presented fairly so that specific stakeholders are not held

fully accountable for problems when multiple factors are involved. Human development is challenging, and the report recognises that implementation teams and national partners are often facing multiple challenges.

6.11. The role of context in initiative performance is analysed

The report identifies relevant aspects of the context within which activities are implemented. These might include geographic, cultural, 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 initiative may have had any effect on the context.

6.12. The text uses appropriate methods/language to convince the reader of the findings and conclusions

Arguments presented do not use emotive word choices in an effort to appeal to the emotions of the reader. The method used to convince readers is the presentation of evidence or a credible basis for the finding. Using the international literature to build the credibility of the report can be effective. The report handles political issues with sensitivity. A good report considers the expected positions of the important stakeholders—if findings are unexpected then this is carefully communicated and explained in the text.

6.13. There is an adequate exploration of the factors that have influenced the issues identified and conclusions drawn

It is not sufficient to simply describe a situation. A full analysis of the likely factors that have led to the situation is necessary. Factors that enable progress or achievement are just as important as factors that inhibit them. These factors should be generated from a range of data sources. A range of causes should be considered rather than regularly offering a single cause for major and/or complex issues.

6.14. The implications of key findings are fully explored

DFAT aid initiative managers, senior management and other stakeholders need some direction on the implications of the findings if this is not immediately apparent. Implications to achieving initiative objectives, implementation for meeting time frames, expenditure projections, or sustainability are often important considerations.

Conclusions and recommendations

6.15. The overall position of the author is clear and their professional judgments are unambiguous

The task of the evaluator is to evaluate. They must make their position clear (and as early as possible in the report) unless the TORs have required the evaluator to report on findings with neutrality. The report does not simply state findings and expect DFAT to interpret them and draw their own conclusions. The report presents the authors view unambiguously. Has the initiative made adequate progress or not? Are the factors that have accounted for the limited achievements been unavoidable or are they due to poor management. Unambiguous judgements also present findings and conclusions sensitively and constructively.

6.16. 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. There are no important inferred recommendations buried in the text that have not been drawn into the conclusion or list of recommendations at the end.

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.

6.17. Individuals have been allocated responsibility for responding to recommendations

Where appropriate, job titles, rather than organisations, have been allocated responsibility for all recommendations for action. If it is not appropriate or possible to identify the individual, then the relevant work group is identified. If some recommendations are for broader partner government, or DFAT sectoral or corporate learning then these are identified separately.

6.18. Significant cost implications of recommendations have been estimated

If recommendations imply human, financial or material costs, these are estimated. If recommendations for additional technical support are made, then the number of days input is estimated. For important technical assistance positions proposed, the key content to consider for the terms of reference is annexed.

6.19 The recommendations are feasible

Recommendations, in the most part, are acceptable to relevant stakeholders (recommendations that stakeholders do not agree with rarely get implemented—coming to acceptability is dealt with by the utilisation strategy). Recommendations are feasible from a resourcing and cost perspective. Recommendations are likely to be effective to rectify a situation, or to achieve an expected outcome.

Annex 6: Good-practice examples

Partnership for Knowledge-based Poverty Reduction in Indonesia

The Partnership for Knowledge-based Poverty Reduction program (PKPR) (INJ244) is a program to support the Government of Indonesia in making informed and evidence-based policy and program decisions. The program has been operating since July 2010 but builds on earlier work that the World Bank was doing in this field. The objectives of this mid-term review were: to evaluate the extent to which AusAID funding has enabled the program to achieve its objectives, identify and synthesise lessons that will drive improvement and review the program's relevance to Government of Indonesia needs and priorities.

Terms of reference

The terms of reference for this evaluation are clear and well structured. The background and the rationale of the evaluation are set out, including recent developments and the purpose is clearly stated as: accountability, program improvements and learning.

It is clear to the consultant what is expected because the scope, duration, process and deliverables of the assignment are clearly defined. The people days and the team composition match the scope of the evaluation, prioritised evaluation questions are provided and evaluation methods mentioned. The evaluation process is mapped out, key documents listed and the roles and responsibilities of the team members are clearly described.

There are only two minor weaknesses in these terms of reference. First, there is an inconsistency regarding the input days for the assignment, with 30 days being suggested in total but 54 days being allocated to the two team members. Second, there is only very limited and implicit mention of the roles and responsibilities of AusAID staff in the evaluation.

Evaluation plan

The evaluation plan is excellent in terms of both quality and coverage. All key elements of an evaluation plan are addressed and presented in a way that is very easy to read, presenting each topic under headings such as 'What is the focus and scope of the review?' or 'Who is the audience?'. It is clear what is proposed to be done, why, how, when, by whom and for whom, and numerous further details are presented.

The evaluation design is well thought through and includes a good justification of the focus, approach and methods used. Different methods are proposed for each of the three main objectives of the evaluation and specified in detail. Key evaluation questions are developed and linked to methods in a well presented evaluation matrix. Different parameters to triangulation are proposed and limitations and constraints are discussed.

The evaluation plan also specifies the primary users of the evaluation and the dissemination mechanisms that will be used. Reference is also made to the codes of ethics the evaluation team will adhere to and examples of how this will be applied in practice are cited.

One area where the plan could be improved is in its discussion of sampling. It only states that 'high-interest stakeholders' will be engaged by the review team without discussing who these stakeholders are or the sampling approach.

Evaluation report

The evaluation report itself is of high quality. It is clear and well structured. It is detailed, but remains easy to read. The report uses the evaluation questions and sub-questions to structure the analysis, which provides the reader with good guidance and ensures that all questions are answered. Both operational and strategic issues are covered in the report.

There is a clear line of evidence between the data, findings and recommendations. There are a few instances where findings are not clearly substantiated by the evidence, but this does not apply to any of the major findings of the report. Equally, the recommendations are well grounded in the evidence and are clear, targeted and actionable. Notably the report prioritises recommendations and discusses resource implications.

The report provides a well-thought-through and comprehensive analysis of the relevance of the initiative, which is also one of the key evaluation questions set out in the terms of reference. Various dimensions of relevance are looked at and numerical rankings assigned and aggregated. While the report does not explicitly refer to effectiveness, a number of evaluation questions relate to the effectiveness of the program and are assessed in a comprehensive manner. Context and its impact on performance is factored into the assessment. Finally, the report presents a very detailed assessment of the program's M&E system and clearly identifies strengths and weaknesses. Notably, the M&E system is rated against the AusAID/DFAT Aid M&E standards, and existing M&E data is used to make a judgment on progress to date.

Nevertheless, the report displays some weaknesses. There is limited analysis of the contribution of the program to policy change and the role played by other factors. Furthermore, the assessment of efficiency and, in particular, the analysis of value for money is weak. Lastly, no assessment of the criteria of impact is provided, despite being requested in the terms of reference.

Timor-Leste Asian Development Bank Infrastructure Project Management/Infrastructure Technical Assistance program

The Timor-Leste Asian Development Bank Infrastructure Project Management/Infrastructure Technical Assistance program (INH497) aims to create and upgrade infrastructure assets in line with the Government of Timor-Leste's medium-term targets, including transport, communications, urban development, power and water supply and sanitation. The program was approved by the ADB in June 2007 and finishes in 2014. The main purpose of this completion report is to identify lessons to inform the Australian Government's ongoing and future programs in the infrastructure sector in Timor-Leste and to respond to the aid program's accountability requirements.

Terms of reference

The terms of reference for this evaluation are very clear and focused. The policy and context of Australia's involvement in the infrastructure sector in Timor-Leste are set out and a comprehensive

overview of the program is provided. Notably, the primary and secondary purpose of the evaluation are defined and specified: program improvement and accountability respectively. Furthermore, primary and secondary users of the evaluation are identified: the Australian aid Timor-Leste infrastructure team in Dili and Canberra and Australian aid Timor-Leste broader program teams in Dili and Canberra and ADB as primary users; and partners in Timor-Leste, including the Ministry of Infrastructure, the International Labour Organization team responsible for the Roads for Development program, and the Australian aid program's Infrastructure Thematic Group as secondary users.

Most importantly, the terms of reference provide a very focused and limited number (six) of clear evaluation questions with sub-questions. These relate to the key objectives of the evaluation and focus on assessing program performance and evaluating the management decisions the Australian aid program has to take to meet its accountability requirements. The scope of the evaluation questions is also well aligned with the number of days allocated for the task. Two team members are expected to deliver the assignment in 34.5 days, excluding travel days. The terms of reference provide clear definitions of the reporting requirements and deliverables of the evaluation and a useful list of the stakeholders that should be engaged through the evaluation and the documents to review.

One minor flaw remains: the terms of reference provide only limited information on the roles and responsibilities of the evaluation team or Australian aid program officers.

Rural Water Supply and Sanitation Program in Timor-Leste

The first phase of the Australia Timor-Leste Rural Water Supply and Sanitation Program (ING002)—known locally as 'Bee, Saneamentu no Ijene iha Komunidade' or BESIK—was implemented between 2007 and 2012. It supported capacity development within key ministries and other sector partners to deliver improved water and sanitation to communities through technical advisors, budget support, organisational development, policy development, training, construction of water systems in rural districts, and research. The primary purpose of this independent completion report was to support program improvement and to inform both the next phase of the program and wider initiatives in the Australian Government's Timor-Leste program.

Evaluation plan

The evaluation plan is comprehensive and detailed. It sets out the background of the program and the evaluation, the evaluation's purpose, and primary and secondary users and the evaluation team. Responding to the terms of reference, the evaluation approach is focused on the efficacy of gender mainstreaming methods and capacity development. A gender outcomes framework is presented and framework for conceptualising and assessing capacity building at the individual, organisational and system level is outlined.

A question guide frames the issues defined in the terms of reference as hypotheses and develops evaluation questions to test them. An evaluation matrix then links these issues and questions to data collection methods and specific evaluation activities. This provides a comprehensive and well-thought-through evaluation framework.

The evaluation plan also discusses sampling, triangulation and limitations. For instance, a sample frame of interviewees is presented and there is a good discussion of the challenge of finding an appropriate balance between logistical pragmatism and methodological rigour.

A number of ethical issues are addressed such as gender sensitivity, the presence of Australian Government/BESIK staff in stakeholder interviews, and how the findings of the evaluation can be

meaningfully fed back to key stakeholders. The Evaluation Plan also presents an indicative outline of the report and an evaluation schedule.

Civil Society Water, Sanitation and Hygiene Fund

The Civil Society, Water, Sanitation and Hygiene Fund (INI592) was implemented by 11 civil society organisations (CSOs) between 2009 and 2011 throughout Africa, Asia and the Pacific. Its goal was to improve the health and quality of life of the poor and vulnerable by improving their access to safe water, better sanitation and hygiene. The main objectives of this completion review were to judge the performance of the fund and to identify lessons for the upcoming Civil Society WASH Fund. Secondary objectives were to review key innovative elements and to describe the value for money delivered by the fund. The review was undertaken by the fund's three-member monitoring review panel, with the exception of one component—an evaluation of the monitoring review panel itself—which was undertaken by the former AusAID's Quality, Performance and Results branch.

Evaluation report

This evaluation report provides an example of how to provide a robust and credible evaluation of a complex global program primarily through a desk-based study. However, it should be noted that this was probably only possible as a result of the evaluators being part of the fund's monitoring review panel and therefore having an in-depth and detailed knowledge of the initiative. The report is well structured and provides a clear orientation for different audiences, identifying clearly which parts of the report are relevant to AusAID/DFAT's Infrastructure Water Policy section, NGO section or country and regional programs, and to partner CSOs.

The report is methodologically robust with a good mix of quantitative analysis based on CSO data and qualitative data. The limitations of this type of desk review are clearly spelled out. Since the review was conducted by members of the fund's monitoring review panel itself, an additional separate evaluation of the monitoring review panel was conducted by the former AusAID's Quality, Performance and Results Section. This evaluation also draws on a survey of CSO partners' perceptions of the monitoring review panel model, which adds to the findings and ensures the independence of the overall review process.

There is a very clear line of evidence from the data to the findings. The recommendations flow naturally from the analysis and are highlighted clearly at the end of each section. They are clear, targeted, relevant and actionable and offer a strong foundation for any future similar fund for CSOs in this sector.

The standard Australian aid quality evaluation criteria are well assessed. Notably the report provides a whole section devoted to value for money with an excellent discussion that includes global value for money data. Good evidence is provided regarding sustainability, as this issue is a major theme of the report. The assessment of gender equality also stands out; unusually, it reflects on the role of men.

A minor flaw is that the report does not sufficiently discuss the objectives of the fund or embed them in a contextual analysis.

Two remote service delivery and community development programs in Papua New Guinea

The Kokoda Development Program (KDP) (INH843) and the Integrated Community Development Program (ICDP) (INJ153) both aim to improve economic opportunities, livelihoods and basic services for remote populations of Papua New Guinea. KDP focuses directly on improving service delivery while ICDP aims at building the capacity of government and civil society to deliver and advocate for services. The objective of the combined mid-term review was to inform decisions around ongoing support to these initiatives and to future programming where remote service delivery may be required.

Evaluation report

This is a high-quality evaluation report that provides a clear structure, a strong line of evidence, and clear recommendations that build on a good analysis of the theory of change of both initiatives.

The report is structured around the aid quality criteria of effectiveness, efficiency and sustainability and analyses both the contributing factors and barriers to each. The sustainability section is particularly strong, with a comprehensive analysis of factors influencing sustainability. Based on this, a two-pronged approach to improve sustainability outcomes is proposed. The report also provides a good analysis of the effects that operating in extremely remote environments have on performance. At the end of each section is a set of clear recommendations directed to specific stakeholders including the program managers and AusAID. The structuring of the report also means that recommendations are grouped according to criteria.

The brief methodological section of the report is thorough and suggests a good balance between different evaluation methods. Limitations and ways of addressing them are discussed. The findings flow logically from the evidence and analysis. Documentation is well referenced and there is good use of case studies to illustrate some of the program achievements with real-life examples.

The report provides in-depth analysis of the intervention logic of both programs and how it compares in practice. An excellent critique of the logframes of both programs is presented. Among other points, it finds that both programs are not sufficiently results focused but report mostly on inputs and outputs. The final section provides a comparative analysis of the different implementation modalities of the two programs and draws useful lessons.

A minor flaw is that the report does not provide any financial analysis that reviews the total costs of the programs against the very small populations that are benefiting. Furthermore, there is no detailed analysis of the stand-alone, isolated nature of the KDP program as a key issue affecting performance. The report would have benefited from factoring these elements into the analysis of efficiency and effectiveness.

Annex 7: List of evaluations reviewed

Evaluations that have been published externally can be accessed through the DFAT aid publications webpage[29] or, in some cases, through aid country program webpages.

Initiative number	Evaluation title	Country	Primary sector	Evidence and analysis assessed as credible?	Published on DFAT website?
INC357	Independent Evaluation of Agusan del Sur Malaria Control and Prevention Program	Philippines	Health	Yes	Yes
IND982	Evaluation of the outcomes and sustainability of the Laos–Australia Basic Education Project (LABEP)	Laos	Education	Yes	No
INE114	Independent Completion Review of Provision of Core Funding Support to the SMERU Research Institute	Indonesia	Improved government	Yes	Yes
INE887 INJ788	Independent Completion Review of Australian Civil Society Program Fiji	Fiji	Human rights	Yes	No
INF725	Evaluation of Africa–Australian Development Scholarships Management Program	Africa	Education	Yes	No
INF759	Independent assessment report and recommendations on possible future activities for Papua New Guinea Media Development Initiative 2	Papua New Guinea	Human rights	Yes	Yes
ING002	Independent evaluation of AusAID's support to rural WASH in Timor-Leste through the Rural Water Supply and Sanitation Program (RWSSP/BESIK)	Timor-Leste	Water and sanitation	Yes	Yes
ING236	South Asia Regional Program Evaluation (AusAID-ADB South Asia Development Partnership Facility and AusAID-World Bank Facility for Decentralisation, Local Governance and Service Delivery)	Multicountry	General development support	Yes	Yes

[29] The DFAT aid publications webpage is currently located at: http://aid.dfat.gov.au/Publications/Pages/List.aspx?publicationcategory=Evaluation%20Reports.

Initiative number	Evaluation title	Country	Primary sector	Evidence and analysis assessed as credible?	Published on DFAT website?
ING357	Evaluation of Public Sector Capability Development Program (PSCDP) in Timor-Leste	Timor-Leste	Improved government	Yes	Yes
ING406	Independent Progress Review of Eastern Indonesia Road Improvement Program	Indonesia	Infrastructure	Yes	Yes
ING661	Final Evaluation Report for Rakhine Rural Household Livelihood Security Project (RRHLSP)	Myanmar	Rural development and food security	Yes	No
ING723	ODE Evaluation of Australian Law and Justice Assistance: Cambodia case study	Cambodia	Security and justice	Yes	Yes
ING754	Mid Term Review of Cambodian Agricultural Value Chain (CAVAC) program	Cambodia	Rural development and food security	Yes	Yes
ING854	Independent Review of the Pacific Technical Assistance Mechanism (PACTAM)	Multicountry	Improved government	Yes	Yes
ING918	End of Program Review of the Papua New Guinea–Australia Sexual Health Improvement Program (PASHIP)	Papua New Guinea	Health	Yes	Yes
ING948	Mid-Term Review of the State- and Peace-Building Fund (SPF)	Multicountry	Conflict prevention and resolution	Yes	No
ING967	Mid-term Review of Implementation Support Program to P135 Phase II in Quang Ngai Province	Vietnam	Rural development and food security	Yes	No
ING982	Independent Completion Report for Regional Rights Resource Team (RRRT)	Multicountry	Security and justice	Yes	No
ING997	Final Evaluation of the Three Diseases Fund	Myanmar	Health	Yes	Yes
INH095	Independent Completion Report for Laos–Australian Scholarships Program	Laos	Education	Yes	No
INH274	Independent Completion Report for Challenging the Frontiers of Poverty Reduction (CFPR) Phase II	Bangladesh	Rural development and food security	Yes	No

Initiative number	Evaluation title	Country	Primary sector	Evidence and analysis assessed as credible?	Published on DFAT website?
INH361	Independent Progress Review of Support for Education Sector Development in Aceh (SEDIA)	Indonesia	Education	Yes	No
INH436	Independent Progress Review of Australia-UNICEF Education Assistance to Papua and Papua Barat	Indonesia	Education	Yes	No
INH497	Independent Completion Report for Timor–Leste Asian Development Bank Infrastructure Project Management/ Infrastructure Technical Assistance	Timor-Leste	Infrastructure	Yes	Yes
INH528	Independent Progress Report for Pacific Leadership Program	Pacific	Improved government	Yes	Yes
INH602	Mid Term Evaluation of the School Sector Reform Program	Nepal	Education	Yes	Yes
INH843 INI153	Independent Review of two remote service delivery and community development programs in Papua New Guinea	Papua New Guinea	Environment and natural resource management	Yes	No
INH157	Independent Progress Report for ASEAN Australia Development Cooperation Program Phase II	Multicountry	Improved government	Yes	Yes
INH947	Evaluation of Education for Children in Areas Affected by Armed Conflict—Mindanao Philippines	Philippines	Education	Yes	No
INI035	External review of the Indonesia Project	Indonesia	Education	Yes	No
INI171	Independent Progress Report for Provincial Road Management Facility (PRMF)	Philippines	Infrastructure	Yes	Yes
INI194	Independent Progress Report for PNG–Australia Law & Justice Partnership	Papua New Guinea	Security and justice	Yes	Yes
INI311	Mid Term Review of the Vanuatu Kastom Governance Partnership Program	Vanuatu	Improved government	Yes	No
INI355	Independent Progress Report for Local Governance Innovations for Communities in Aceh, Phase II (LOGICA2)	Indonesia	Improved government	Yes	Yes

Initiative number	Evaluation title	Country	Primary sector	Evidence and analysis assessed as credible?	Published on DFAT website?
INI422	Independent Progress Review of the Australia Indonesia Facility for Disaster Reduction	Indonesia	Humanitarian response	Yes	No
INI426	Independent Evaluation of Lessons Learned from UN Delivering as One	Multicountry	General development support	Yes	No
INI510	Review of the Afghanistan Reconstruction Trust Fund	Afghanistan	General development support	Yes	Yes
INI592	Independent Completion Review of the Civil Society Water, Sanitation and Hygiene Fund	Multicountry	Water and sanitation	Yes	Yes
INI632	Mid Term Review of the Australia–WB Philippines Development Trust Fund	Philippines	General development support	Yes	Yes
INI661	Independent Review of the SIG-RAMSI Public Sector Improvement Program (PSIP) Year 4	Solomon Islands	Improved government	Yes	No
INI674	Independent Progress Review of the Solomon Islands Media Assistance Scheme Phase 3	Solomon Islands	Human rights	Yes	No
INI865	Independent Mid Term Review of the Australian Community Rehabilitation Program Phase 3	Sri Lanka	Humanitarian response	Yes	Yes
INI903	Independent Progress Review of the PNG–Australia Economic and Public Sector Program (EPSP)	Papua New Guinea	Improved government	Yes	No
INI941	Mid Term Review of AusAID NGO Cooperation Program Partnership Agreements	Multicountry	Human rights	Yes	Yes
INJ052	Independent Review of the Tonga Sector Consolidation Project (TSCP)	Tonga	Infrastructure	Yes	No
INJ124	Independent Completion Report for the Padang Pariaman Health Facility Reconstruction Program	Indonesia	Health	Yes	Yes
INJ135	Interim Review of the Livelihoods and Food Security Trust Fund	Myanmar	Rural development and food security	Yes	Yes
INJ152	Independent Progress Report for Solomon Islands Clean Water & Sanitation Program	Solomon Islands	Water and sanitation	Yes	No

Initiative number	Evaluation title	Country	Primary sector	Evidence and analysis assessed as credible?	Published on DFAT website?
INJ189	Independent Completion Report for Zimbabwe NGO Food and Water Initiative	Zimbabwe	Water and sanitation	Yes	No
INJ197	Final Annual Performance Assessment 201130 for Kiribati Technical Vocational Education and Training Sector Strengthening Program Phase I	Kiribati	Business, finance and trade	Yes	Yes
INJ235	Independent Progress Review of Education in Emergencies Capacity Building	Multicountry	Education	Yes	No
INJ241 INJ398 INI309	Independent Review of Two AusAID Funded UNICEF Projects on Child Survival and Nutrition and Maternal Health in Nepal	Nepal	Health	Yes	Yes
INJ244	Independent Progress Review of Partnership for Knowledge-Based Poverty Reduction (PKPR)	Indonesia	Improved government	Yes	Yes
INJ251 ING400	Independent Progress Review of the UN Joint Program on Maternal and Neonatal Mortality Reduction	Philippines	Health	Yes	Yes
INJ321	Independent Review of AusAID's Support to the UN in PNG through the UN Country Fund	Papua New Guinea	General development support	Yes	Yes
INJ344	Independent Review of Australia Africa Community Grants Scheme	Multicountry	General development support	Yes	No
INJ371	Final Evaluation of Timor-Leste Investment Budget Execution Support for Rural Infrastructure Development and Employment Generation	Timor-Leste	Infrastructure	Yes	Yes
INJ632	Independent Progress Review of the Australia Indonesia Electoral Support Program 2011–15	Indonesia	Human rights	Yes	No
INJ657	Mid Term Review of Partnership agreement between AusAID and Australian Red Cross	Multicountry	Humanitarian response	Yes	Yes
INJ675	Independent Progress Report for the Vanuatu Australia Police Project	Vanuatu	Security and justice	Yes	No

[30] Completed in 2012.

Initiative number	Evaluation title	Country	Primary sector	Evidence and analysis assessed as credible?	Published on DFAT website?
INJ691	Mid Term Independent Review of the RedR Australia and AusAID Partnership Agreement	Multicountry	Humanitarian response	Yes	Yes
INJ746	Independent External Review of the Secretariat of the Pacific Community (SPC)	Multicountry	General development support	Yes	Yes
INJ794	Independent Review of MTV Exit Asia III	Multicountry	Human rights	Yes	Yes
INK299	Review of the Pacific Islands Forum Secretariat (PIFS)	Multicountry	Improved government	Yes	No
INK557	Mid Term Review of Strengthening Food Security for Rural Livelihood Program	Solomon Islands	Rural development and food security	No	No
INJ526 INJ194	Independent Completion Report for Burma Program WASH Activities	Myanmar	Water and sanitation	No	No
INI568	Independent Completion Report for Prevention and Control of Avian and Human Pandemic Influenza in Myanmar (Phase II)	Myanmar	Health	No	No
INH284	Review of the Australia Awards Program, Cambodia	Cambodia	Education	No	Yes
INI660	Mid Term Review of Financial and Economic Management Strengthening Program	Solomon Islands	Governance	No	No
ING833	Focused evaluation of Micro Enterprise Development Program	Nepal	Business, finance and trade	No	Yes
ING953	Independent Review of Tingim Laip Phase II	Papua New Guinea	Health	No	Yes
INJ828	Mid Term Evaluation of Support to Conflict-Affected People through Housing in Sri Lanka	Sri Lanka	Humanitarian response	No	No
INI767	Mid Term Review of Strongim Gavman Program	Papua New Guinea	Improved government	No	Yes
INI954	Independent Progress Review of Tonga Technical and Vocational Education and Training (TVET) Support Program	Tonga	Education	No	Yes

Initiative number	Evaluation title	Country	Primary sector	Evidence and analysis assessed as credible?	Published on DFAT website?
INJ137	Independent Progress Review of Australia Indonesia Partnership for Justice	Indonesia	Security and justice	No	Yes
INI220	Evaluation of Australia's Support to Agriculture in Papua New Guinea	Papua New Guinea	Food security and rural development	No	No
INH487	Independent Completion Review of AusAID Timor-Leste Justice Sector Support Facility	Timor-Leste	Security and justice	No	Yes
INI651	Mid Term Review of Mekong River Commission Integrated Capacity Building Program 2009–11	Multicountry	Water and sanitation	No	No
IND653	Independent Completion Review of the Centre for Democratic Institutions	Multicountry	Human rights	No	No
ING522	Independent Completion Report for Australia China Environment Development Partnership	China	Environment and natural resource management	No	No
INH459	Final Evaluation of Youth Employment Promotion Programme	Timor-Leste	Education	No	No
INJ823	Independent Progress Review of Australia's Support to the Government of Indonesia Tim Bantuan Tata Kelola Pemerintahan	Indonesia	Improved government	No	No
INI661	Independent Progress Reprot for Solomon Islands Electoral System	Solomon Islands	Human rights	No	No
INH074	Mid Term Review of Vois Blong Yumi—Vanuatu Media Strengthening	Vanuatu	Human rights	No	No
INI691 INH466	Independent Evaluation of the Infrastructure Partnerships Program and the Water and Sanitation Initiative Global Program	Multicountry	Water and sanitation	No	Yes
INI307	Mid-Term Review of the Vanuatu Church Partnership Program	Vanuatu	Human rights	No	No
INI988	Independent Review of the Commonwealth Local Government Forum Good Practice Scheme—Phase II	Papua New Guinea	Improved government	No	No