Annual Assessment of Results (AAR)

Report of the AAR for the New Zealand Aid Programme in the Ministry of Foreign Affairs and Trade for AMAs and ACAs completed in 2017/18

Final Report

Prepared for //

Insights, Monitoring & Evaluation
Pacific and Development Group
New Zealand Aid Programme
Ministry of Foreign Affairs and Trade

Date // 3 July 2019

IOD PARC is the trading name of International Organisation Development Ltd//

Omega Court
362 Cemetery Road
Sheffield
S11 8FT
United Kingdom

Tel: +44 (0) 114 267 3620
www.iodparc.com

Contents

Acronyms

ACA	Activity Completion Assessment
AMA	Activity Monitoring Assessment
AQP	Activity Quality Policy
ARF	Activity Results Framework
CCIs	Cross Cutting Issues
DAC	Development Assistance Committee
DP&R	Development Planning and Results (team)
M&E	Monitoring and Evaluation
MFAT	Ministry of Foreign Affairs and Trade
MO	Multilateral Organisation
N/A	Not Assessable
ODA	Official Development Assistance
RMT	Results Measurement Table
SAM	Scholarships and Alumni Management System
UN	United Nations
VfM	Value for Money

Executive Summary

Background

New Zealand's Aid Programme is delivered through investments (known as Activities) administered by the Ministry of Foreign Affairs and Trade (MFAT). The performance of individual Activities is reported through Activity Monitoring Assessments (AMAs) and Activity Completion Assessments (ACAs). Within AMAs and ACAs, Activity Managers rate performance on a five-point scale against set criteria. Upon completion of an Activity, ACAs record ratings of effectiveness, relevance, impact, efficiency, and sustainability. AMAs record ratings of effectiveness of on-going Activities. An important element of AMAs and ACAs is that they identify issues that may affect the implementation and results of activities, as well as recommendations and lessons to improve activities.

For many Activities, AMAs and ACAs are the only formal MFAT assessment of their progress and performance. It is therefore important that MFAT has confidence in their robustness. The Annual Assessment of Results (AARs) assesses the robustness of ratings in AMAs and ACAs. The AAR also assesses whether AMAs and ACAs are helpful to improve Activities

According to Activity Managers who were interviewed for this AAR, AMA and ACA templates have become more user-friendly over the last five years through revisions based on outcomes of AARs, amongst others. AMAs and ACAs are becoming increasingly embedded within MFAT's performance management system. Completion rates for AMAs and ACAs have increased substantially from 59% in 2013/14 AAR to 88% in 2017/18, the highest completion rate to date.

This is the fourth (AAR)[1], providing an independent quality assurance of a randomly selected representative sample of 66 AMAs and 36 ACAs, drawn from a total of 141 AMAs and 55 ACAs that were completed between 1 July 2017 and 30 June 2018. Of those selected, 63 AMAs and 28 ACAs were eventually reviewed.[2] AMAs and ACAs with non-robust effectiveness ratings were progressed to an interview with the relevant Activity Manager, following which reviewers finalised their assessment.

A proportionally low number of interviews conducted with Activity Managers to review non-robust effectiveness ratings is a major limitation of this AAR[3]. To enhance comparability with previous AARs, final effectiveness ratings were adjusted to allow for the low interviewing rate in the current AAR.[4]

Summary of Findings

Robustness of Effectiveness Ratings

To have general confidence in the robustness of Activity Managers' effectiveness ratings across all AMAs and ACAs, MFAT would expect at least 75% of the assessed ratings to be robust.

MFAT can be reasonably confident in the robustness of AMA effectiveness ratings, especially for short- and medium-term outcome ratings. Based on the adjusted post-interview ratings, it is encouraging that the robustness of short-term outcome ratings in **AMAs** meets the 75% confidence threshold, while that of medium-term outcome ratings is just under, at 74%. The robustness of output ratings still trails at 64%.

[1] Three previous AARs were conducted in 2015, 2016 and 2018 for AMAs and ACAs completed between 1 July 2013 – 30 June 2014; between 1 July 2014 – 30 June 2015; and between 1 July 2016 – 30 June 2017, respectively.

[2] The main reason for the differences between selected and reviewed totals is that Activity-related documents could not be provided in time to be included in the review.

[3] The interview rate for the current AAR (33%) is much lower compared to previous AARs (86% in the inaugural AAR; then 92% and 84% in the two subsequent AARs).

[4] The adjustment for each result area is based on the average percentage change in post-interview robustness of effectiveness ratings across the previous three AARs and applying it to the pre-interview rating in the current AAR. Both the initial post-interview robustness ratings and the adjusted post-interview ratings have limitations. The initial post-interview ratings are not informed by a proportional number of interviews compared to previous years and therefore are likely to be understated. The adjusted post-interview ratings are indicative rather than definitive, but are likely to be closer to what the actual results may have been.

The robustness of output and medium-term outcomes ratings for AMAs in the current AAR is lower compared to previous AARs, but the robustness of short-term outcome ratings is higher. However, there has been no statistically significant changes in the robustness of AMA effectiveness ratings in the current AAR compared to both the inaugural AAR and the 2016-17 AAR.[5]

Based on adjusted post-interview ratings, the robustness of short- and medium-term outcome ratings in **ACAs**, at 77%, exceeds MFAT's confidence threshold, but the robustness of output ratings trails at 72% - compared to the inaugural AAR, the decrease in robustness of ACA output ratings is likely to be statistically significant[6]. When an Activity ends, it is especially important to know whether it has achieved its results or not. It is therefore encouraging that MFAT can be reasonably confident about the robustness of short- and medium-term outcome ratings in ACAs.

In both AMAs and ACAs, effectiveness ratings of 1 and 2 (inadequate ratings) are more robust compared to ratings of 4 and 5 (good and very good ratings). This suggest that Activity Managers may be inclined to over-rate effectiveness of Activities, which is consistent with all previous AARs.

As in previous AARs, robustness of effectiveness ratings for both AMAs and ACAs increased substantially after interviewing. The main reason is that Activity Managers may not document all the evidence that informs ratings in the AMA or ACA. Contextual information and understanding are major factors that influence Activity Managers' ratings, but this is often not explained in AMAs and ACAs. Activity Managers may assess an Activity's progress and performance with due consideration of challenges in the implementing context, but do not always document this in AMAs and ACAs. Therefore, many AMAs and ACAs still do not capture the evidence that supports effectiveness ratings in sufficient detail to provide stand-alone records of an Activity's effectiveness.

Quality of Results Management Tables and influence on Effectiveness Ratings

Based on references to Results Management Tables (RMTs) as sources of evidence for completing AMAs and ACAs, Activity Managers are increasingly relying on RMTs to monitor and report Activities' progress. Ratings were often substantiated by references to movements in RMT indictors, or progress in relation to baselines and targets. An increasing number of Activity Managers are also proposing appropriate actions to address identified shortcomings of RMTs.

However, RMT shortcomings continue to impede the quality of AMAs and ACAs. In the current AAR, 15% of AMAs (n=9) were based on RMTs assessed as adequate, compared to 68% (n=40) where RMTs were found to have shortcomings. Many RMTs are not updated regularly and may not reflect the evolving reality of dynamic Activities. Other major shortcomings of RMTs include the absence of baselines, targets and data to monitor and report progress. It appears that completing AMAs and ACAs for complex Activities, for example multi-county and multi-donor Activities, as well as Activities funded through Budget support, could be challenging because RMTs for these Activities are not straight-forward. Per MFAT guidelines, AMAs for funding to Multilateral Organisations draw on the Strategic Plans and annual reports of the organisations themselves. Their annual reporting is therefore not based on an MFAT RMT and does not provide a clear-cut fit with AMA assessment criteria.

Robustness of other DAC Criteria ratings in ACAs

MFAT guidelines for the assessment of the DAC criteria of relevance, efficiency, impact and sustainability were revised in 2017 and incorporated in AMA and ACA templates. However, the reviewers based their assessment of the robustness of ratings on outdated guidelines in Annex 1 of the 2015 Aid Quality Policy (AQP). With this limitation in mind, the robustness of DAC criteria ratings in ACAs has predominantly improved compared to previous AARs but remains below the 75% confidence threshold. More specifically:

[5] Assessment of statistical significance is limited by availability of detailed data from the inaugural AAR and the adjustment approach applied in 2017/18.
[6] Assessment of statistical significance is limited by availability of detailed data from the inaugural AAR and the adjustment approach applied in 2017/18.

- the robustness of relevance, efficiency and sustainability ratings has improved, following almost consistent decreases over the previous three AARs.
- the robustness of impact ratings is higher compared to the first two AARs but has decreased slightly compared to the previous AAR. A key challenge in assessing an Activity's impact at completion appears to be the absence of baseline data.

Bearing in mind that that ratings of these criteria were not discussed during interviews, and despite encouraging increases in robustness of sustainability and impact ratings, confidence in these ratings would still be somewhat lacking.

Cross-Cutting Issues

While the assessment of Cross-Cutting Issues (CCIs) remains a challenging aspect of AMAs and ACAs, it is encouraging that the proportion of ACAs containing 'adequate' or better analyses of CCIs has increased substantially over time. However, the proportion of CCI analyses in AMAs that received 'good' or 'very good' ratings remains small and has decreased since the previous AAR. This suggests that analyses provided by Activity Managers may still lack evidence, and/or depth and insight, to warrant high ratings.

Actions to Address Issues (AMA) and Lessons Learned (ACAs)

In AMAs, the overall quality of proposed actions to enhance Activities' performance has increased over successive review periods. Compared to the inaugural AAR, the proportion of AMAs that identified lessons assessed as inadequate decreased by 13%. At the same time, the proportion of lessons assessed as 'good' or 'very good' has also decreased slightly (8% from baseline), suggesting that actions to strengthen on-going Activities are pooling in the 'adequate' category.

In ACAs, lessons to inform future Activities remain of a relatively high quality – 82% of ACAs identified meaningful, useful lessons to strengthen future activities. This is 19% more compared to the inaugural AAR, but 14% less that the 2016-17 AAR. It is very encouraging that the proportion of ACAs that did not identify any lessons, or identified generic, unsubstantiated lessons, has decreased by 51% compared to the inaugural AAR.

Drawing on AMAs and ACAs to improve Activities

The extent to which AMAs and ACAs can be used to improve Activities considers whether they articulate a plausible, evidence-based story of an Activity's progress and performance, and then identify key issues and helpful recommendations/lessons to strengthen the Activity (or Activities in general) going forward. The assessment is based on different criteria to the effectiveness ratings and therefore results may diverge.

Despite some missed opportunities to identify actions or lessons that could improve Activities, MFAT can be cautiously confident that both AMAs and ACAs can be drawn on to improve Activities. Qualitative elements in the majority of AMAs and ACAs provide insightful assessments across a range of issues, and they identify meaningful actions and lessons to inform activity improvement.

- The proportion of AMAs and ACAs that cannot be drawn on to improve activities has decreased steadily, but so has the proportion that are considered highly informative. This has resulted in a net increase the proportion that are considered 'adequate' in documenting information that can be drawn on to improve Activities. Therefore, most AMAs and ACAs contain helpful information that can be drawn on to improve Activities.
- A relatively large proportion of AMAs (20%, or six AMAs) for Activities with large whole-of-life costs ($5 million and more) did not provide helpful information to improve the associated Activities. This can be ascribed partly to the fact that most complex Activities, as well as funding of Multilateral Organisations, have large budgets but there is limited scope for proposing ways within MFAT's control to strengthen them.

Conclusions

1. Despite improvements in qualitative aspects of reporting, AMAs and ACAs still do not provide sufficiently comprehensive or stand-alone records of Activities' progress. As in previous AARs, Activity Managers draw on evidence from a range of sources to assess the effectiveness of Activities but tend not to document all this evidence in AMAs and ACAs. If the evidence is not comprehensively documented, the loss of institutional knowledge leaves substantial gaps, especially where staff turnover is high. When these gaps build up year-on-year, new Activity Managers might find it challenging to complete insightful AMAs and ACAs, thereby jeopardising the robustness of AMAs and ACAs in the longer term.
2. Despite remaining challenges around the robustness of effectiveness ratings, AMAs and ACAs generally include helpful information to improve Activities. Providing helpful information to improve complex Activities, which often have high whole-of-life costs, are challenging since ways to improve these Activities may not be within MFAT's full control.
3. So far, AARs have not revealed major statistically significant results related to AMA and ACA improvement over time. However, due to contextual factors such as the capacity and capability of Activity Managers, incentives, etc. in an environment of relatively high staff turnover, MFAT does not expect to see statistically significant linear improvements over time.

Recommendations

1. AMAs and ACAs should remain as essential building blocks of the Aid Programme's performance management system. Activity Managers use AMAs and ACAs to reflect and assess the progress, performance and challenges of Activities and they serve as important repositories of institutional memory and continuity during Activity implementation. Increasingly insightful and usable lessons and actions to address issues, if harnessed through a robust knowledge management system, could also prove valuable in strengthening Activities.
2. Ongoing training and technical support would be important to ensure that gains made in the robustness and usefulness of AMAs and ACAs are maintained and enhanced. Gradual improvements are becoming evident, but it would be important to address known challenges and strengthen capacity to maintain this positive momentum and to prevent the gains made from being lost.

   Continue to provide training and guidance for Activity Managers in Wellington and at Post (including locally-engaged staff) to ensure that they understand why and how to document the evidence base for AMAs and ACAs fully, yet concisely, to increase the proportion of AMAs and ACAs that provide stand-alone records of Activities' progress and performance. This would be instrumental to lift the robustness and usefulness of AMAs and ACAs (and therefore their value as essential building blocks of the Aid Programmes performance management system) to a higher level.

   Training and support in the following priority areas could be considered:

   - Documenting consolidated evidence from several sources to justify effectiveness and DAC criteria ratings.
   - RMT quality and wider socialisation of RMTs as foundations of Activity design, monitoring and reporting, as well as dynamic tools for Activity improvement.
   - MERL expert assistance to support regular reviewing and updating of RMTs to ensure that they remain relevant and up-to-date.
   - Strengthening RMTs and monitoring of complicated and complex Activities, for example multi-donor and multi-country Activities, as well as Activities funded through budget support.
   - Improving consistency and coherence in AMAs and ACAs, including identifying issues that affect the progress and performance of Activities, and following this through into

meaningful, evidence-based actions to improve on-going activities (in AMAs), or lessons relevant to comparable types of Activities when they complete ACAs.

3. Given the size of MFAT's funding to Multilateral Organisations and the unique arrangements around their monitoring, it could be beneficial to tailor guidance for the AMAs of these Activities.
4. Provide support to Activity Managers to identify and perceptively address appropriate Activity cross-cutting markers:
   - Where a cross-cutting marker is identified as relevant, it should be dealt with consistently and perceptively throughout the design, monitoring and reporting of the activity, including in AMAs and ACAs.
   - Avoid including cross-cutting markers that are not relevant to an Activity in its AMA/ACA.
5. AARs should continue to be conducted on a periodic basis to monitor the effect of known enablers and constraints to the robustness and usefulness of AMAs and ACAs, as well as to identify emerging challenges and actions for their continuous improvement. A larger database will also enable meaningful trend analyses of the robustness and usefulness of AMAs/ACAs across different sectors, programmes and budget levels.

1. Introduction

Background

New Zealand's Overseas Development Assistance (ODA) is delivered through Activities administered by the Ministry of Foreign Affairs and Trade (MFAT). AMAs and ACAs are internal assessments of the Activities' performance and provide forward-looking actions and lessons to improve Activities. They form the building blocks of the Aid Programme's Performance System.

- AMAs are completed annually for Activities expending over $250,000 per annum, or smaller Activities with a high-risk profile. They rate and describe effectiveness of on-going Activities, and provide a descriptive assessment of performance against relevance, efficiency, sustainability, cross-cutting issues, risk, Activity management and actions to address identified issues.
- On completion of activities with a total expenditure over $500,000 ACAs are completed, ideally within one month of receiving the final Activity Completion Report from the implementing partner. ACAs rate and provide a narrative assessment of an Activity's performance in relation to five Development Assistance Committee (DAC) criteria, namely relevance, effectiveness, efficiency, impact and sustainability; they also comment and provide analyses on cross-cutting issues, risk and Activity management, and identify lessons to improve future Activities.

Completion rates for AMAs and ACAs have increased since 2013/14. The completion rate for 2014/15 AMAs and ACAs was 79%, which was an increase of almost 20% from 2013/14. The completion rate for 2016/17 AMAs and ACAs was 76%. AMAs and ACAs from 2017/18 had the highest completion rate ever, namely 88%.

During interviewing, eight (of 11) Activity Managers noted that revisions of AMA and ACA templates over the last five years have made their completion easier, more useful and less time consuming. Three Activity Managers stated that they had completed training and knew where to access support if they required assistance with the completion of AMAs/ACAs. Two Activity Managers – one based at post and one in New Zealand – noted that some staff at Post were unaware/unable to access the same level of training/support that is offered to Wellington-based staff. Both stated that this could be improved by increased promotion of available support at Post and including the information in orientation training for locally-engaged staff, in particular.

Purpose of the AAR

For many Activities, AMAs and ACAs completed by Activity Managers are the only formal MFAT assessment of their progress and performance. Aggregated ratings data from AMAs and ACAs provide a snapshot of the performance of New Zealand's ODA. It is important to have confidence in the robustness of these ratings. AARs assess the robustness of Activity Managers' ratings of Activities' effectiveness. It also assesses the overall usefulness of AMAs and ACAs, as well as the analysis of cross-cutting issues, actions proposed to address issues (AMAs), and of lessons learnt (ACAs) presented by Activity Managers.

The purpose of the AAR is to:

- assess the level of confidence that MFAT can have in the robustness of AMAs and ACAs
- inform the Insights, Monitoring & Evaluation team's efforts to strengthen the Aid Programme's Performance System
- provide input to the Aid Programme Strategic Results Framework.[7]

[7] Level 3 of the New Zealand Aid Programme Strategic Results Framework incorporates the following indicators pertaining to the quality of AMAs and ACAs:
2.3: Percentage of AMAs and ACAs rated 3 or higher on a scale of 1-5 reviewed against quality standards
2.4: Percentage of AMAs and ACAs rated 3 or higher on a scale of 1-5 reviewed against quality standards for Cross-cutting issues

This is the fourth Annual Assessment of Results (AAR) that provides an independent quality assurance of AMAs and ACAs. It was conducted for AMAs and ACAs completed between July 2017 and July 2018.[8] It describes findings from a sample of 66 AMAs and 36 ACAs, drawn from a total of 141 AMAs and 55 ACAs. Of those selected, 63 AMAs and 28 ACAs were eventually reviewed.[9] Effectiveness ratings in 34 AMAs and 18 ACAs were found to be inconsistent with documented evidence and 35 Activity Managers[10] were identified for interviewing to clarify these ratings. Interviews were conducted with 11 Activity Managers, covering eleven AMAs and six ACAs. This constitutes an overall interview rate of 33%, which is much lower compared to previous AARs, and which is a major limitation of the current AAR.

Structure of the Report

The report starts with a brief overview of the AAR's purpose (Section 1) and methodology, including limitations (Section 2). Section 3.1 deals with the robustness of effectiveness ratings in AMAs and ACAs, while Section 3.2 deals with the robustness of other rated criteria, namely relevance, efficiency, sustainability and impact (ACAs only). Challenges to the robustness of effectiveness ratings are also discussed, including the role of Results Management Tables (RMTs). Section 3.3 reflects on the assessed quality of non-rated criteria, namely Actions to address Issues Identified (AMAs), Lessons Learned (ACAs) and the overall usefulness of AMAs and ACAs. The analysis of Cross-Cutting Issues is also discussed. The report ends with conclusions and recommendations in Sections 4 and 5 respectively.

2. Methodology

The methodology for the AAR was developed in consultation with MFAT and has been refined over four successive AARs. A summary is provided here, with more detail in Appendix 1.

Sampling

The AAR is based on a statistically representative sample of AMAs and ACAs that met the following criteria:

- it was submitted between July 2017 and July 2018
- it was completed within 12 months after the assessment period ended.

A total population of 141 AMAs and 55 ACAs met these criteria. A simple random sample of 66 AMAs and 36 ACAs (95% confidence levels, 10% confidence intervals). Ultimately, the AMA sample was reduced to 63 (that is 45% of AMAs) due to unavailability of information. To ensure consistency with previous AARs, AMAs for Scholarship Activities were excluded from all comparative analyses, meaning that comparative analyses are based on an AMA sample of 59.[11] The ACA sample was reduced to 28 (51% of ACAs) for the same reason. The distribution of the samples according to sectors, programme categories and Whole-of-Life Budget Programme Approvals is illustrated in Appendix 3.

There were no major changes to the quality reporting system during the review period. The use of revised AMA and ACA templates, which were introduced in June 2017, is well-established. Most AMAs (91% of AMAs) and ACAs (75% of ACAs) were completed in the revised templates. Due to the small number of AMAs and ACAs that were completed in the old templates, an analysis of the effect of template revisions on the robustness of AMAs and ACAs would not be meaningful.[12]

[8] The previous three AARs were the inaugural AAR, conducted in 2015 for AMAs and ACAs completed between July 2013 and July 2014; the AAR conducted in 2016 for AMAs and ACAs completed between July 2014 and July 2015, and the AAR conducted in 2018 for AMAs and ACAs completed between July 2016 and July 2017.

[9] The main reason for the differences between selected and reviewed totals is that Activity-related documents could not be provided in time to be included in the review.

[10] Some Activity Managers were responsible for more than one AMA or ACA.

[11] The Scholarships Programme completed AMAs for the first time in 2014/15. Four Scholarship Activities were included in the 2016 AAR and eight were included in the previous AAR. The four scholarship activities included in the current AAR sample account for 3% of the AMA sample.

[12] Five AMAs and seven ACAs were complete using the old templates.

As in previous AARs, AMAs for Scholarship Activities are excluded from the main analysis because the robustness of scholarships AMAs have typically been significantly lower and have skewed overall AMA robustness results. In previous AARs, AMAs were prepared for individual Scholarship Activities. In the current AAR period, the approach to AMAs for the scholarship Activities have been revised to have a few integrated AMAs focused at a more programmatic level. Therefore, there was an over-arching AMA prepared for the Scholarship programme, as well as integrated AMAs for all Scholarship Activities at a country level. There was consistency of assessments and issues across the four integrated Scholarship AMAs that were reviewed. Therefore, compared to previous AARs, the influence of scholarship AMAs on the robustness of effectiveness ratings in the current AAR would have been less evident, but for consistency scholarships activities are excluded.

Assessments

Standard assessment templates for AMAs and ACAs were developed for the inaugural AAR and slightly amended over previous AARs without compromising comparability with the inaugural AAR. For the current AAR, two revisions were made to templates:

- The rating scale for assessing RMTs in the AMA template was re-organised so that ratings reflect a logical increase in quality from a rating of 1 to a rating of 5.
- In previous AARs, an overall assessment of AMAs and ACAs focused on quality, based mainly on the robustness of evidence that informed the completion of the AMA/ACA. In the current AAR, the overall assessment considered the robustness of evidence, report coherence and the extent to which this evidence was used to propose helpful actions or lessons to improve Activities. It considers coherence between identified issues and proposed actions/lessons to address these issues.

A desk review of the sampled AMAs and ACAs, as well as accompanying partner progress reports, completion reports and the reports of independent evaluations, was conducted to assess the robustness of effectiveness ratings and, in ACAs, also the ratings for other DAC criteria. The assessment of robustness is based on the extent to which Activity Managers' ratings were substantiated by the evidence and analysis presented in the AMA/ACA, as well as in supporting documentation provided to reviewers, in accordance with guidance in the Activity Quality Policy (AQP) and in the new AMA and ACA templates. To have confidence in the robustness of Activity Managers' effectiveness ratings across all AMAs and ACAs, MFAT would expect at least 75% of the assessed ratings to be robust.

Non-rated elements of AMAs and ACAs were subject to more qualitative assessments, based on tailored rating scales. These assessments are not subject to the 75% confident level.

Ratings scales are in Appendix 1, while copies of the assessment templates are in Appendix 5.

Interviews

Effectiveness ratings in 34 AMAs (52%) and 18 ACAs (50%) were initially assessed as non-robust. Ideally, Activity Managers responsible for all these AMAs and ACAs should have been interviewed telephonically to clarify the evidence base for the effectiveness ratings. However, only eleven Activity Managers, covering only one-third of the identified AMAs and ACAs, were interviewed.[13] Following interviews, a final assessment of the robustness of ratings was made.

Interviewing resulted in an increase in the robustness of all effectiveness ratings for both AMAs and ACAs (see Table 1). Compared to before interviewing, the percentage of outputs assessed as robust after interviewing increased by 7% for AMAs and 7% for ACAs. After interviewing, the robustness of both short-term and medium-term ratings in AMAs increased by 5%, while the robustness of outcome ratings in ACAs increased by 7%.

Where non-robust ratings were revised to robust after interviewing, Activity Managers could provide additional information or explanations to justify the ratings they had given. This evidence is not systematically documented in AMAs or ACAs, for example judgements formed from project visits and direct interaction with implementing partners and/or partner governments. Contextual

[13] Of the 11 interviews that were conducted, four involved discussion of more than one AMA and/or ACA.

information and understanding are major factors that influence Activity Managers' ratings, but this is often not explained in AMAs and ACAs. Activity Managers may assess an Activity's progress and performance with due consideration of challenges in the implementing context, but do not always document this in AMAs and ACAs.

The difference between pre- and post-interviewing robustness of effectiveness ratings in AMAs and ACAs can be seen in the third and fourth columns of Table 1 (Robust effectiveness ratings before interviews and Initial robust effectiveness ratings after interviews).

Where interviewing did not result in the revision of ratings from non-robust to robust (on average, 51% of AMA and 42% of ACA robustness ratings changed), the Activity Managers could not provide evidence to justify the robustness of the rating they had given. The reasons for this include lack of available evidence from progress reports, inability to triangulate evidence from more than one source, and/or inability to conduct field visits.

The increase in effectiveness ratings robustness through interviewing indicates that AMAs and ACAs cannot currently be considered stand-alone records of an Activity's progress and performance, which consistent with the finding of previous AARs.

Adjusted Effectiveness Ratings

Compared to previous AARs, a much smaller number of Activity Manager were interviewed as part of the current AAR. Due to scheduling challenges, only 11 out of 35 Activity Managers were interviewed, covering six out of 18 ACAs and eleven out of 34 AMAs; giving an average interview rate of 33%. In previous AARs, interviewing resulted in notable increases in the robustness of effectiveness ratings, but interview rates were much higher: 86% in the inaugural AAR; and 92% and 84% in the two subsequent AARs, respectively.

To allow for the low interview rate in the current AAR compared to previous AARs, adjusted robustness rates were calculated. These are based on the average percentage change in robustness of effectiveness ratings post-interview across the previous three AARs, which was then applied to pre-interview effectiveness ratings in the current AAR as a multiplier. For AMAs, these average percentage increases came to 45%, 42% and 24% for output, short-term outcome and medium-term outcome AMA ratings, respectively. For ACAs, it came to 34% and 20% for output and short- and medium-term effectiveness ratings, respectively. When applied as a multiplier to respective pre-interview ratings, these increases are reflected as Adjusted effectiveness ratings (see Table 1).

Regardless of the adjustment made to the effectiveness ratings, the pre-interview ratings indicate a divergence in the Activity Manager's rating and the assessed rating either because of lack of evidence to justify the rating, or a discrepancy between the applicable guidance and the provided evidence.

Table 1: Percentage of robust effectiveness ratings in AMAs and ACAs before and after interviews, based on actual and adjusted interview rates (excludes AMAs for Scholarship Activities)

	Effectiveness Criteria	Robust effectiveness ratings before interviews	Initial robust effectiveness ratings after interviews*	Adjusted effectiveness ratings**
AMAs (n = 59)	Outputs	44%	51%	65%
	Short-term Outcomes	53%	58%	75%
	Medium-term outcomes	59%	64%	74%
ACAs (n = 28)	Outputs	54%	61%	72%
	Short and Medium-term outcomes	64%	71%	77%

* These figures are based on a relatively low interview rate of 33%, compared with interview rates of 86% in the inaugural AAR; and 92% and 84% in the two subsequent AARs, respectively. ** Adjusted effectiveness ratings are based on an average percentage increase multiplier, see Appendix 1.

Limitations of the AAR

1. The AAR is based on selected information on Activity performance. This includes AMAs, ACAs, the partner report(s) corresponding to the completion dates of AMAs/ACAs, as well as reports of independent evaluations, where available. Interviews with Activity Managers supplement the information base for some assessments. Other information that may affect ratings is not reviewed.
2. Comparisons of actual post-interview effectiveness ratings between the current and previous AARs should be made with caution.

   Both the initial post-interview robustness ratings and the adjusted post-interview ratings have limitations. The initial post-interview ratings are not informed by a proportional number of interviews compared to previous years and therefore likely presents and under-estimation of robustness of effectiveness ratings. The adjusted post-interview ratings are indicative rather than definitive, but are likely to be closer to what the actual results may have been had a similar proportion of interviews compared to previous AARs been conducted.
3. The assessment of DAC criteria (including effectiveness) was based on Quality Criteria Considerations outlined in Appendix one of MFAT's 2015 Aid Quality Policy. However, Activity Managers likely based their ratings on the AMA/ACA Guidelines revised in 2017. This includes simplified requirements for Effectiveness, Relevance, Impact, Sustainability and Efficiency, and Cross-Cutting issues being addressed in the Effectiveness section rather than throughout all DAC criteria. This could affect the comparability of assessed ratings in the current AAR with those of previous AARs.
4. The sample size is a representative sample determined at 95% confidence level and a confidence interval of 10. Decreasing the confidence interval will increase the sample size and reduce the margin of error. This will enable more precise estimates of AAR findings.[14] However, for the 2017-18 AAR, confidence intervals have only been calculated for the initial post-interview ratings and not the adjusted post interview ratings due to potential limitations in the approach.
5. Keeping interviews to 30 minutes each means that they focus on non-robust effectiveness ratings, as well as four process-related questions. Other non-robust ratings (e.g. for other DAC criteria in ACAs), or the analyses of CCIs, Lessons and Actions can generally not be discussed within the available time.
6. The AAR was carried out by two assessors, and pairs of assessors have also changed in subsequent AARs. Inter-assessor consistency is therefore a potential limitation. This was mitigated through (1) In-depth orientation and induction of assessors by an experienced Team Leader, who has been involved in all four AARs to date; (2) Assessor 'calibration' following the assessment of two AMAs and one ACA; (3) Cross-checks and investigation of inconsistent findings compared to previous AARs; and (4) Independent quality assurance of selected AMAs and ACAs completed by both assessors.

[14] Taking initial outputs ratings in Annex 1 as an example, this means that we can say with 95% certainty that between 41% and 60% of AMAs will have effective output ratings, no matter how many different samples of the same size we draw form the total number of AMAs. Another way of saying it is that 51% of AMAs have robust output ratings, with a 19% margin of error. If the sample size is increased, the confidence interval (and margin of error) will become smaller.

3. Key findings

3.1 Robustness of effectiveness ratings

For AMAs, Activity Managers rate effectiveness of Activities separately for outputs, short-term outcomes and medium-term outcomes. For ACAs, the effectiveness of Activities is rated for outputs, while a combined rating is given for short- and medium-term outcomes.

The current AAR included an integrated AMA for multiple scholarship Activities, comprising scholarships for New Zealand based tertiary, Commonwealth, Regional Development Scholarships and In-Country English Language Training. It incorporated some, but not all the AMAs in the other three selected scholarship Activities, namely scholarships for Commonwealth, Regional Development, Pacific Development and Short-Term Training in Niue, Timor-Leste and Tuvalu, respectively. Compared to the previous AAR, the influence of scholarship AMAs on the robustness of effectiveness ratings in the current AAR has been similar, although less prominent. When AMAs for scholarship Activities were excluded from analyses, the robustness of output and short-term outcome ratings increased, while that of medium-term outcomes decreased, but not as markedly as in the previous AAR.

The quantitative analysis of results is based on the final, adjusted, post-interview assessment of the robustness of ratings in AMAs and ACAs, excluding AMAs for Scholarship Activities. To be confident in the robustness of effectiveness ratings, MFAT would expect at least 75% of these ratings to be robust.

Robustness of effectiveness ratings: AMAs

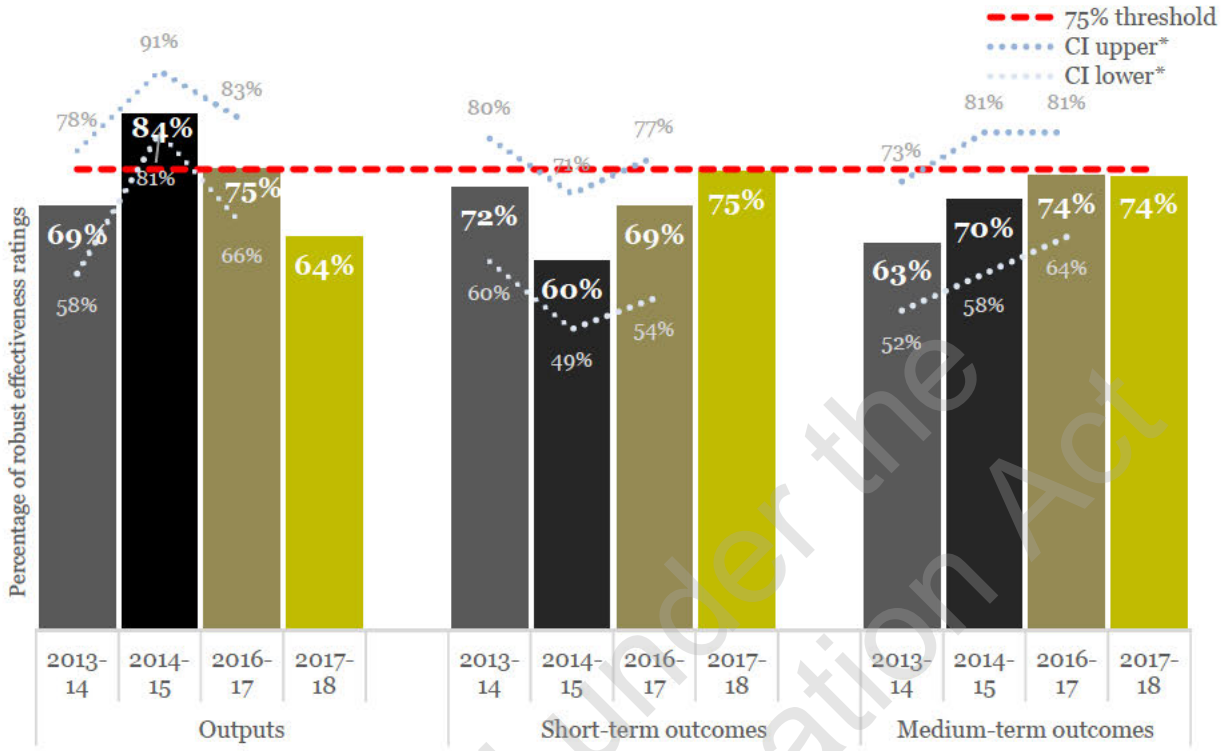
The robustness of output ratings in the current AAR remains below the 75% confidence threshold and is also lower compared to previous AARs (Figure 1). The robustness of medium-term outcome ratings would be just below the 75% threshold – higher than the baseline and similar to the previous AAR. The robustness of short-term outcome ratings would be higher compared to the previous three AARs and, for the first time, it would meet the 75% threshold.

Assessment of likely statistical significance between the adjusted 2017-18 results and both the inaugural AAR and the 2016-17 results indicated no statistically significant changes. [15]

This suggests that MFAT can be reasonably confident in the robustness of AMA effectiveness ratings.

[15] Assessment of statistical significance is limited by availability of detailed data from the inaugural AAR and the adjustment of ratings applied in 2017-18.

Figure 1: Robustness of AMA effectiveness ratings across four AARs, based on adjusted ratings for the current AAR
(AMAs for Scholarship Activities excluded)



*Due to the adjustment of post-interview ratings in the current AAR, confidence intervals could not be calculated for FY17-18.

Quality of Activity Results Management Tables (AMAs only)

Since 2011, each Activity funded by the New Zealand Aid Programme has been required to have an Activity Results Framework (ARF), including a Results Measurement Table (RMT) which sets out indicators, baselines and targets to track progress towards intended results. AMAs and ACAs assess and report an Activity's progress and performance against the results outlined in its RMT. The inaugural AAR found that weaknesses of RMTs had a negative effect on the quality of AMAs and ACAs. More in-depth analysis of RMTs has been undertaken as part of subsequent AARs to gain a better understanding of their strengths and weaknesses.

The rating scale for assessing the quality of RMTs was revised for the current AAR, so direct comparisons with findings from previous AARs are not always possible (see ratings scales, Annex 1).

RMTs in a total of 59 AMAs were reviewed (the AMAs for the four scholarship Activities were not assessed). Key findings pertaining to the quality of RMTs pre-interview can be summarised as follows:

- The number of Activities that had no RMT decreased from 31% in the second AAR to 8% (n=5) in the previous AAR, and to 2% (n=1) in the current AAR.
- In the current AAR, 15% of AMAs (n=9) were based on RMTs assessed as adequate, compared to 68% (n=40) where RMTs were found to have shortcomings[16]. 15% of AMA RMTs (n=9) were 'not-assessable'. They could not be reviewed because they were not included in the AMAs or partner reports, or in the case of funding for Multilateral Organisations, where effectiveness is monitored and reported by the organisations themselves against their strategic plans.
- In the previous AAR, the RMTs of 81% (n=49) of sampled AMAs had RMTs with shortcomings, but only 19 Activity Mangers (39%) identified and addressed shortcomings of the RMT as an

[16] One AMA did not have an RMT, while 9 RMTs could not be assessed.

impediment to completing AMAs. In the current AAR, it is encouraging that Activity Managers made appropriate recommendations to strengthen RMTs in 68% (n=27) of the 40 AMAs that were found to have shortcomings.

- None of the RMTs reviewed were found to be a 'good practice' example.

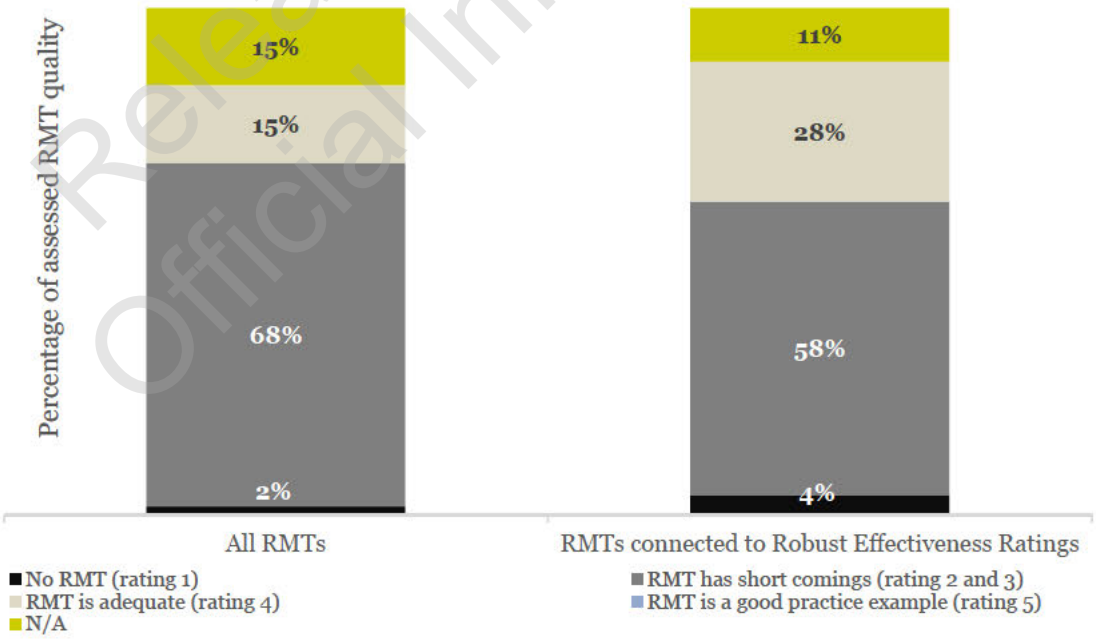
Compared to previous AARs, robust ratings in the current AAR appear to demonstrate greater confidence in evidence on the part of Activity Managers. Ratings were often substantiated by references to movements in indictors, or progress in relation to baselines and targets. The reviewers also noted that, where Activity Managers included little or no narrative to substantiate ratings in an AMA/ACA itself, they could often find evidence in RMTs to assess the robustness of ratings.

Box 1
Monitoring MFAT's funding for Multilateral Organisations

AMAs for these Activities draw on M&E frameworks and standard reporting by the relevant Multilateral Organisations. It could therefore be challenging to distinguish outputs and outcomes, and to source the evidence required for robust ratings. Also, where progress is not deemed satisfactory, MFAT may have little influence to address it. Some Activity Managers called into question the purpose of monitoring long-standing MFAT aid commitments to Multilateral Organisations.

The relationship between the quality of RMTs and robustness of effectiveness ratings is illustrated in Figure 2. Compared to the robustness of all effectiveness ratings, the proportion of robust effectiveness ratings in AMAs based on RMTs assessed as adequate is almost twice larger. Fifteen percent of all effectiveness ratings in AMAs are based on RMTs assessed as not adequate, whereas 28% of robust effectiveness ratings are based on RMTs assessed as adequate. Where RMTs have shortcomings, the robustness of effectiveness ratings decreases by 10% compared to the robustness of all effectiveness ratings. Further information can be found in the section on challenges associated with the robustness of effectiveness ratings.

Figure 2: Relationship between assessed quality of RMTs and robustness of effectiveness ratings in AMAs (excluding Scholarships AMAs)



*** No RMT received a quality rating of 5**

Key shortcomings identified through an assessment of 59 AMAs for RMT quality are summarised in Table 2.

Table 2: Shortcomings of RMTs in AMAs - comparison of the previous and current AARs

	Number and % of RMTs in 2016/17 AMA sample with this shortcoming (N=69)*	Number and % of RMTs in 2017/18 AMA sample with this shortcoming (N=59)**
Indicators and targets		
Targets are too broad to be measurable, while indicators and data sources are not realistic.	6 (9%)	6 (10%)
No targets / indicators identified; or there are only some but not for all outputs or outcomes.	6 (9%)	11 (19%)
Targets are set conservatively. Targets are exceeded quite early on in an Activity's lifetime.	4 (6%)	2 (3%)
Targets are over ambitious	0	1 (2%)
There is an over-reliance on quantitative indicators with no supporting qualitative evidence to contextualise or explain changes in quantitative measures.	2 (3%)	0
Changes in methods of data collection have impacted how targets are calculated	1 (1%)	0
It is not clear how indicators relate to baselines and targets.	1 (1%)	5 (8%)
There is repeated use of one indicator that is applied to outputs and outcomes.	1 (1%)	0
Indicators are listed, but they are not linked to specific outputs or outcomes.	1 (1%)	1 (2%)
Data is missing for many indicators.	1 (1%)	6 (10%)
Indicators do not relate to outcomes.	1 (1%)	0
Baseline		
No baseline data is available. (Baseline data are still to be collected and/or completed.)	6 (9%)	14 (24%)
Outputs and outcomes		
Outputs and/or outcomes are not identified, or their quality is questionable.	17 (25%)	15 (25%)
Short or medium-term outcomes are missing or there is no distinction between them	13 (19%)	14 (24%)
Outputs are phrased as outcomes	3 (4%)	1 (2%)
There are too many and/or too complex outcomes	1 (1%)	1 (2%)
Program Logic		
Results pathways or logics are not clear. The relationship between outputs and outcomes is not well defined or not evident.	8 (12%)	3 (5%)
RMT has only one short-term outcome, but several medium and long-term outcomes.	1 (1%)	0
Outputs are driving activities, but there are no outcomes to reflect the Activity's strategic importance for beneficiaries.	1 (1%)	1 (2%)
RMT logic is unclear, and/or there are issues with the presentation and lay out	0	6 (10%)
Data collection and reporting against RMT		
Partner is not effectively reporting against targets, or not using the RMT to report progress. Evidence of progress towards the achievement of results does not align with indicators. The RMT is not sufficiently operationalised and the partner needs support to collect quality data and to report.	11 (16%)	14 (24%)
Indicators are not yet measured for outcomes, or no data is available.	6 (9%)	2 (3%)
No/limited reflection or measurement of outcomes in partner reporting.	5 (7%)	2 (3%)
Cross-cutting priorities		
Cross-cutting issues are not sufficiently incorporated in relevant results and indicators, or they are dealt with in a 'mechanistic' manner. For example, sex-disaggregated indicators are considered to be mainstreaming of gender equity, but not linked to any background information or analysis to contextualise its appropriateness or use.	5 (7%)	2 (3%)
RMT no longer fit for purpose		
Either the RMT is part of bigger activity and is not fit for purpose for MFAT, or needs updating to fit current situation, or was developed for a particular phase of Activity and needs updating for a new phase.	5 (7%)	19 (32%)

* Totals will not add up to N (69) since any particular RMT may have more than one shortcoming.
** Totals will not add up to N (59) since any particular RMT may have more than one shortcoming.

There are some differences in the shortcomings of RMTs found in the current AAR compared to the previous AAR. In the previous AAR, shortcomings of RMTs were mainly related to Outputs and Outcomes, e.g. outputs and outcomes were not clearly identified or distinguished; there were too many, or they were too complex (34 out of 69 RMTs, that is almost half, had such shortcomings). This was followed by challenges around Indicators and Targets (24 out of 69 RMTs, that is 35%), with challenges around data collection and reporting against RMT also being common (22 out of 69 RMTs, that is 32%). In the current AAR, most shortcomings were also related to Indicators and Targets (32 out of 59, that is 54%) and Outputs and Outcomes (31 out of 59, that is 53%). However, compared to the previous AAR, substantially more challenges were associated with out-of-date RMTs and RMTs that were no longer fir-for -purpose, as well as shortcomings related to data collection and reporting. There has also been an increase in the number of RMTs where baseline data were absent.

In the current AAR, the most common shortcomings relate to outdated RMTs, or RMTs that were no longer fit-for-purpose (19 out of 59, that is 33%), followed by poorly identified outputs and/or outcomes (15 out of 50 RMTs, that is 25%, had shortcomings in this regard). In the previous AAR, the most common specific RMT shortcomings were poorly identified outputs and outcomes, and poorly distinguished outputs and outcomes (17 out of 69, or 25%, and 13 out of 69, or 19%, respectively). Additional shortcomings noted in the current AAR that were not found in the previous AAR relate to targets being over-ambitious, lack of overall clarity and inadequate presentation/layout of the RMT, as well as inadequate accuracy of data/monitoring.

It appears that the development and updating of RMTs to ensure their relevance and use as dynamic frameworks for monitoring and strengthening Activities on an on-going basis are becoming increasingly challenging. Also, partner reporting against RMT indicators and targets is also falling short by a substantial margin. However, it is encouraging that an increasing number of Activity Managers are proposing appropriate actions to address the identified shortcomings of RMTs – of the 40 AMAs that had RMTs with identified shortcomings, 67% included appropriate recommendations to revise and/or strengthen the Activity's RMT. This could indicate that more Activity Managers are realising the value of robust RMTs and are keen to ensure they have fit-for-purpose to monitor Activities. More robust RMTs would be more useful as Activity monitoring tools and would likely encourage more coherent partner reporting against expected results.

Training and technical assistance could be instrumental in enhancing the robustness of RMTs. Special attention could be given to the development and socialisation of RMTs as foundations of Activity design, as well as dynamic tools for Activity monitoring, reporting and improvement. Progress reporting for more complex Activities, or for Activities that do not require RMTs per MFAT policy might require special attention.

Robustness of effectiveness ratings: ACAs

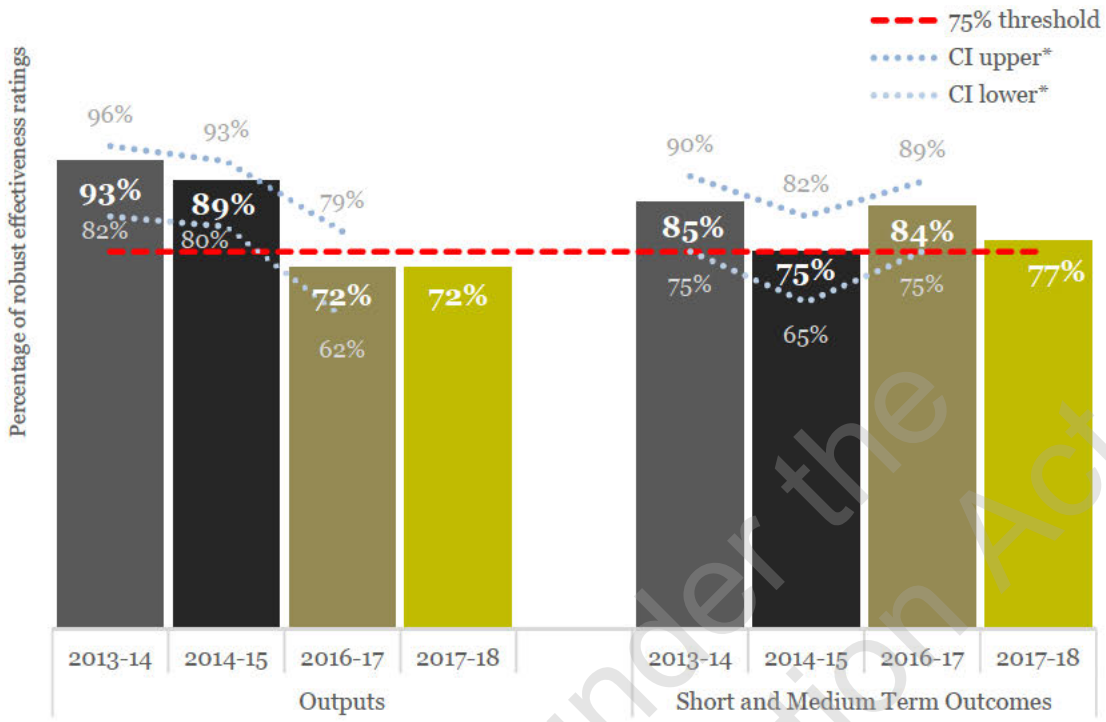
Based on adjusted post-interview effectiveness ratings, the robustness of output ratings is 72% and that of short-and medium-term outcomes is 77% (see Figure 3). This brings the robustness of output ratings in line with the 2016-17 AAR (yet still lower compared to the baseline), while the robustness of short- and medium-term results now exceeds the 75% confidence threshold (but is also still lower compared to the baseline).

Assessment of *likely* statistical significance between the adjusted 2017-18 results and both the inaugural AAR and the 2016-17 results indicates a potentially statistically significant decrease in the robustness of output ratings between the inaugural AAR and the current AAR.[17]

When an Activity ends, it is especially important to know whether it has achieved its results or not. It is therefore encouraging that MFAT can be reasonably confident about the robustness of short- and medium-term outcome ratings in ACAs.

[17] Assessment of statistical significance is limited by availability of detailed data from the inaugural AAR and the adjustment approach applied in 2017-18.

Figure 3: Robustness of ACA effectiveness ratings across four AARs, based on adjusted ratings for the current AAR



* Due to the adjustment of post-interview ratings in the current AAR, confidence intervals could not be calculated for 2017-18.

Robustness of effectiveness ratings across rating levels

Table 3 summarises the robustness of the initial (unadjusted) post-interview effectiveness ratings across different rating levels (1 to 5) in AMAs and ACAs. The unadjusted results are used in this section and in Table 3 because the adjusted results are based on averages, rather than actual findings. This approach does not conflict with other effectiveness findings that use the adjusted rates as the focus is on understanding which rating levels tend to be found to be more robust based on available evidence (both in the AMA/ACA and provided supporting documents).

Most effectiveness ratings across all rating levels were robust. The exception were ratings of 4 and 5 (Good and Very Good) in AMAs, where only 44% of ratings were robust.

In AMAs, 90% of N/A ratings were found to be robust. This means that the reviewer agreed with the Activity Manager that a rating could not be given of the extent to which progress towards the achievement of an Activity's outputs, short-term outcomes and/or medium-term outcomes has been made (also see last paragraph in the section "Factors that affect robustness of effectiveness ratings in AMAs and ACAs").

Ratings of 1 and 2 (inadequate ratings) were proportionately more robust (64% of these lower ratings were found to be robust) compared to ratings of 3 (adequate) at 59%, while ratings of 4 and 5 (Good and Very Good) were least robust (44% of these higher ratings were found to be robust).

In ACAs, reviewers found all inadequate ratings (ratings of 1 and 2) to be robust. This is followed by N/A ratings (86% were found to be robust), adequate ratings (ratings of 3), of which 78% were assessed as robust; and least robust were good and very good ratings (ratings of 4 and 5), at 74%.

This suggests that Activity Mangers may tend to over-rate the effectiveness of both on-going and completed Activities. This finding is consistent with all previous AARs.

Table 3: Robustness of AMA and ACA effectiveness ratings across initial (unadjusted) post-interview ratings

	Number and percentage of robust effectiveness ratings across rating levels							
	Inadequate (Rating 1, 2)		Adequate (Rating 3)		Good and Very Good (Rating 4,5)		No rating (N/A)	
	AMA	ACA	AMA	ACA	AMA	ACA	AMA	ACA
Robust	16	1	20	7	38	23	28	6
	64%	*100%*	*59%*	*78%*	*44%*	*74%*	*90%*	*86%*
Non-robust	9	0	14	2	49	8	3	1
	36%	*0%*	*41%*	*22%*	*56%*	*26%*	*10%*	*14%*

Factors that affect robustness of effectiveness ratings in AMAs and ACAs

Variations in the robustness of effectiveness ratings across AARs can be ascribed to various reasons, amongst others: changes in rating scales over time; the revision of AMA and ACA templates in 2016-17; Activity Managers' access to training and technical assistance; varying AAR interview rates; application of either the Activity Quality Policy 2015 or Activity Quality Ratings 2017 in assessing ratings; differing perspectives between Activity Managers and reviewers on the appropriate effectiveness ratings based on available evidence; and year-on-year variations in the review team.[18]

Further insights into challenges that may affect the robustness of effectiveness ratings in AMAs and ACAs were gained through interviews and include the following:

- Shortcomings of RMTs, as well as challenges and delays in partner reporting relate to the lack of data/evidence to substantiate ratings. For ACAs, the robustness of effectiveness ratings could also be affected by the absence of baseline data.
- Challenges and delays for Wellington-based Activity Managers to obtain data and evidence from Posts could make it difficult for the Activity Managers to assess an Activity's progress.
- Inaccurate interpretation and application of rating scales and criteria on the part of Activity Managers mean that ratings may not be consistent with available evidence.
- It could be challenging for Activity Managers to provide accurate effectiveness ratings where
  - an Activity is generally "difficult to manage", and/or its progress has been hampered by factors that are outside MFAT's or the implementing partner's control (natural disasters, for example);
  - there is a strong element of learning associated with an Activity, for example where it is in a new MFAT investment priority, using an unfamiliar modality or engaging with a new kind of partnership.
- Where achievement of an Activity's medium-term outcomes appears to be likely, Activity Managers may pay less attention to output ratings. This may result in non-robust output ratings.
- Staff turnover in MFAT and partner organisations leads to loss of institutional memory and capacity related to Activities. Each new Activity Manager brings different approaches and perspectives to assessing an Activity's effectiveness, so ratings over time may lack consistency. This reiterates the importance of documenting all relevant evidence that underpins assessments in AMAs and ACAs.

Where effectiveness ratings were assessed as robust and no interview was needed, ratings were supported by:

- Good quality RMTs that enabled the collection and reporting of robust evidence to support ratings. Amongst the 27 AMAs that had robust effectiveness ratings, 24 (89%) were found to

[18] Since the inaugural AAR, leadership of the AAR team in IOD PARC has remained constant, although different reviewers had been involved.

have a robust RMT, which enabled ratings to be substantiated by reporting against clearly identified results and indicators. The same applies to nine out of eleven ACAs (82%).

- Compared to previous AARs, robust ratings in the current AAR appear to demonstrate greater confidence in evidence on the part of Activity Managers. Ratings were often substantiated by references to movements in indictors, or progress in relation to baselines and targets.

In 24 AMAs (excluding scholarship AMAs) and five ACAs, N/A effectiveness ratings were assessed as robust – this means that the reviewers agreed that an Activity's effectiveness could not be rated. Reasons for this included:

- Shortcomings of RMTs. In 17 AMAs (including AMAs for Activities involving Multilateral Organisations and those funded through Budget Support, as well as multi-country and multi-donor Activities) and two ACAs (including one Humanitarian Response Activity), the absence or shortcomings of RMTs made it impossible to assess effectiveness.
- In five of 24 AMAs (21%), it was too early to assess progress towards outcomes.
- In seven of 24 AMAs (29%), there was no data/evidence available to assess effectiveness, including three AMAs where baseline data were not yet available. Effectiveness in four ACAs could not be assessed due to lack of data/evidence.

3.2 Robustness of other DAC criteria ratings (ACAs only)

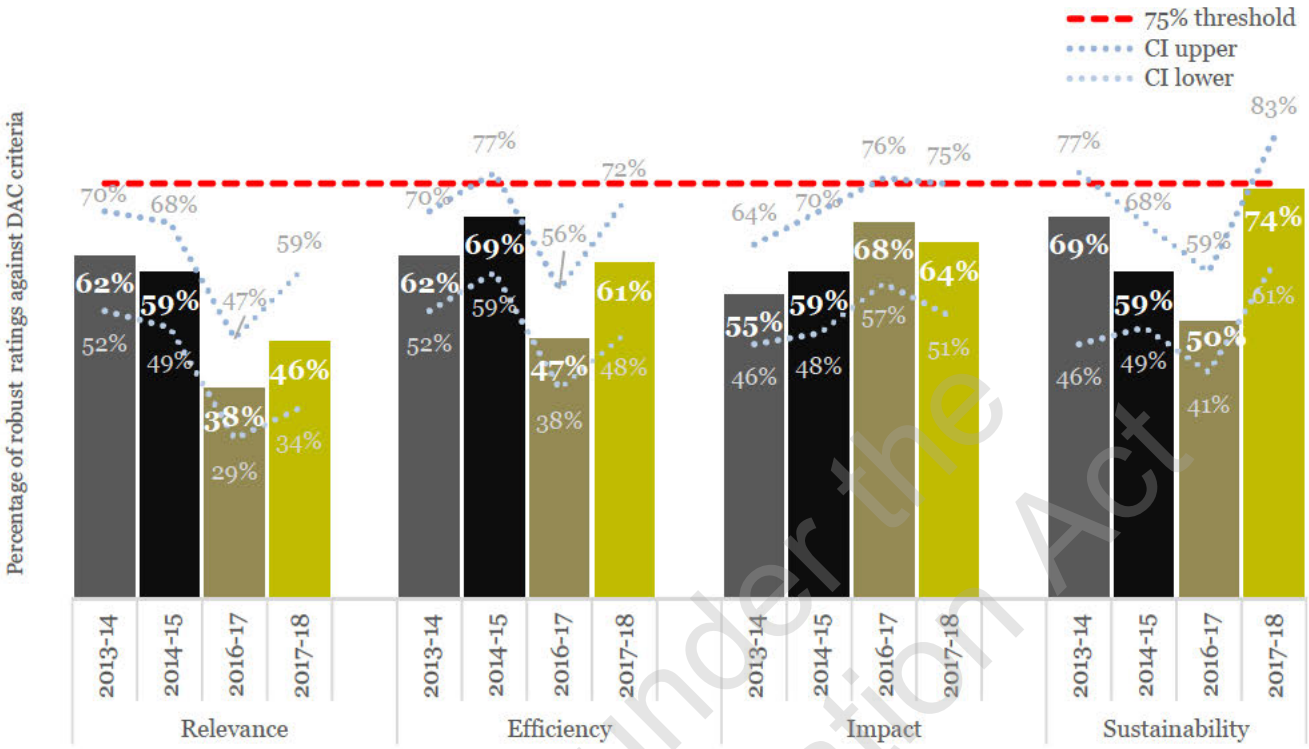
The assessed robustness ratings for the DAC criteria of relevance, efficiency, impact, sustainability in ACAs is based on desk reviews, as interviews focused on effectiveness ratings. Had the DAC ratings been discussed further at interview, there may have been some increase in their robustness in line with increases seen in previous AARs, where the ratings of these criteria were discussed during interviews.

The comparative robustness of ratings for the DAC criteria across the four AARs is shown in Figure 4. Compared to the baseline (inaugural AARs), the robustness of sustainability and impact ratings has increased. While that of efficiency has decreased by only one per cent, the robustness of relevance ratings has decreased by 16%. No rating has yet reached the MFAT confidence level of 75%. Bearing in mind that that ratings of these criteria were not discussed during interviews, and despite encouraging increases in robustness of sustainability and impact ratings, MFAT cannot use the ratings of other DAC criteria in ACAs with confidence to inform decision-making.

The main reason for non-robust relevance ratings is that they do not address all considerations outlined in the 2015 AQP, against which they were assessed. However, MFAT staff may have, appropriately, based their response on the AMA and ACA guideline *Activity Quality Ratings for Completion Assessment* revised in 2017, which is not consistent with the 2015 AQP. Key changes include: simplified requirements for Relevance, Impact, Sustainability and Efficiency, and Cross-Cutting issues being addressed in the Effectiveness section rather than throughout all DAC criteria. Therefore, aspects such as the relevance of modality and the mainstreaming of cross cutting issues across all criteria, are of lesser importance.

A likely result of this would be an increase in the number (and proportion) or robust relevance, efficiency, impact and sustainability ratings – and the robustness of sustainability ratings would likely exceed the 75% confidence threshold. Reviewers based their assessment of the robustness of these ratings on a number of considerations that are no longer included in the revised guidance. Had assessments been based on the revised guidance, more ratings may have been assessed as robust.

Figure 4: Robustness of ratings for other DAC criteria across the four AARs (ACAs only)



3.3 Assessment of qualitative analyses in AMAs and ACAs

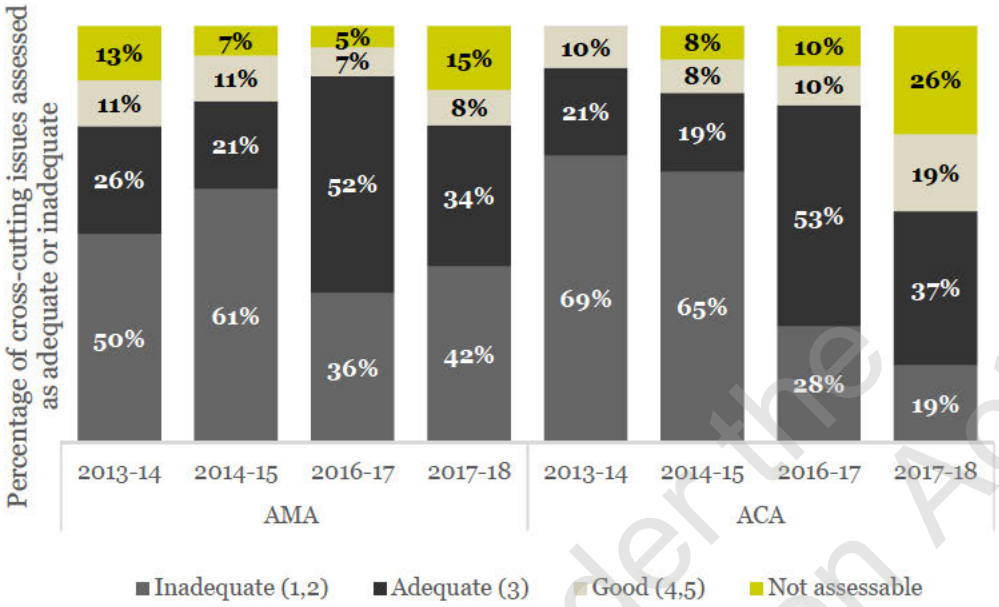
AMAs and ACAs include sections where Activity Managers provide qualitative analyses of CCIs, and where they propose actions to address issues that may be affecting the progress or performance of on-going Activities (in AMAs), or where they identify lessons that could enhance the quality of Activities in the future (in ACAs). The AAR assessed these analyses and, based on their robustness and coherence across all sections of the AAR, also assessed the overall usefulness of AMAs and ACAs (rating scales for the qualitative elements are in Annex 1.

Cross-Cutting Issues

MFAT's cross-cutting priorities include Gender Equality, Human Rights, Environment, Climate Change Adaptation and Climate Change Mitigation. AMAs and ACAs must identify which of these cross-cutting issues (CCIs) are relevant to the associated Activity and must provide an overview of how relevant CCIs are dealt with in the Activity. The AAR assessed the quality of these analyses in accordance with a rating scale that is based on guidance in the AQP (see Appendix 1).

The quality of the analysis of CCIs in ACAs has improved substantially since 2013/14 (Figure 5). An apparent shift from inadequate ratings (ratings of 1 and 2) to 'adequate' ratings (a rating of 3), is also encouraging. However, compared to 2016/17, the quality of CCI analyses in AMAs has decreased.

Figure 5: Assessment of analyses of cross cutting issues in AMAs (Scholarship Activities excluded in AMAs) and ACAs across the four AARs



The proportion of ACAs where the analysis of CCIs was 'good' or 'very good' increased from 10% in the inaugural AAR to 19%, and is the highest result to date. The proportion of ACAs where the analysis was inadequate (ratings of 1 and 2) decreased from 69% to 19%; that is a 50% reduction compared to the inaugural AAR and represents a consistent reduction over time.

The proportion of CCI analyses in AMAs that received 'good' or 'very good' ratings remains small. This suggests that most analyses provided by Activity Managers lacked evidence, and/or depth and insight, to warrant a better rating. However, there appears to be a shift from inadequate ratings (ratings of 1 and 2) to 'adequate' ratings (a rating of 3), which could be encouraging.

The proportion of AMAs and ACAs where CCI analyses received N/A (not assessible) ratings increased and is the highest to date. Reviewers rated cross-cutting issues as N/A where there were no cross-cutting priorities identified in the AMA or ACA (i.e. all cross-cutting markers were classified as 'not measured' or 'not targeted'); or where no or very limited discussion of CCIs were included.

Specific reasons why analyses of CCI were found to be inadequate include the following:

- The AMA/ACA identifies a particular CCI as a targeted cross-cutting marker for an Activity, but offers no analysis for it. This is frequently the case for Human Rights.
- The analysis is based on insufficient evidence and lacks depth or insight. For example, in the case of Gender Equality, quantitative sex-disaggregated data are presented without contextualised analysis relevant to the Activity.

Four interviewees stated that the analysis of CCIs was challenging. The reasons for this include the lack of attention and analysis related to CCIs in the design of Activities; omission of results and/or indicators related to CCIs in RMTs; and shortcomings in data collection to inform monitoring and reporting on CCIs, even when they are included in the design and RMT. It was suggested that capacity building for activity managers on analysis of cross cutting issues would be beneficial as some consider their understanding of these issues as 'superficial' and miss them in the reporting, or just assume they have been included in the activity.

Lessons Learned (ACAs)

The ACA template requires Activity Managers to identify lessons learned from an Activity upon completion, while AMAs require Activity Managers to outline actions to address identified issues that are affecting the progress or performance of an Activity. Rating scales for assessing the analyses are in Appendix 1.

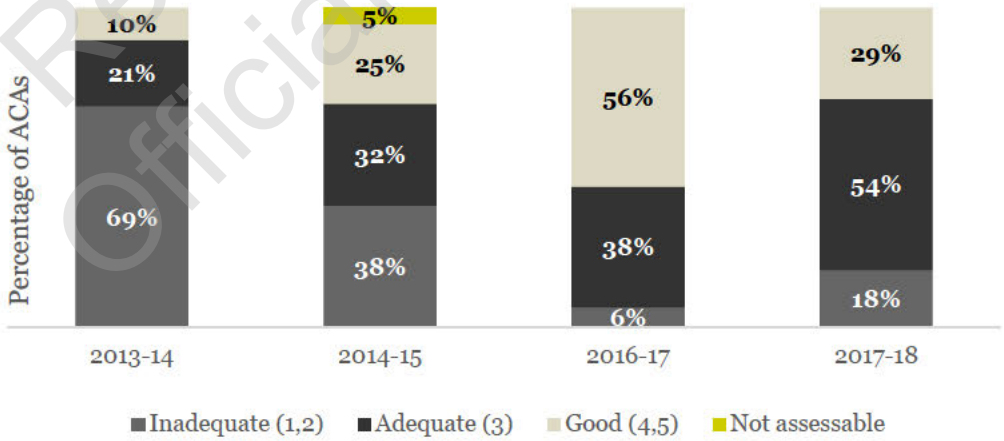
The rating scale for assessing lessons to improve Activities in ACAs makes an important distinction between generic and specific lessons. Generic lessons include lessons that are commonly known as being key to a 'good' Activity, for example establishing a good relationship with the implementing partner, or the importance of monitoring progress. Specific lessons draw on evidence to provide useful insights that could strengthen MFAT's programming in a particular geographical, thematic, cross-cutting or technical area. Compared to the baseline, the proportion of ACAs that identified well-substantiated, useful lessons increased from 10% to 29% - that is an increase of 19% (Figure 6), though far lower than the 56% found in 2016-17. Some ACAs included insightful lessons about specific issues, for example capacity development[19], activity planning[20] and humanitarian support[21].

Over time, there has been a consistent increase in the percentage of ACAs that identify helpful lessons to improve Activities. There has also been a potentially encouraging decrease in the proportion of ACAs that did not identify any lessons, or identified generic, unsubstantiated lessons (at 18%, while an increase on the 2016-17 results, this proportion is a 51% decrease compared to the inaugural AAR and lower than the 38% in the 2014-15 AAR).

Despite these positive changes, the assessment found that Activity Managers do not consolidate lessons consistently. In 13 out of 28 ACAs, the assessment found that the lessons learnt component was a missed opportunity for activity-related learning and MFAT programming. For example, there was a major missed opportunity in the ACA for an Activity that was implemented in two communities to draw comparisons between the communities and to learn from this. This was due to shortcomings of the M&E approach. Although similar data were collected in both communities, the data were not disaggregated to enable comparisons between the two communities.

These lost opportunities indicate that further training and technical support would be required to ensure that Activity Managers identify meaningful, evidence-based lessons relevant to comparable types of Activities when they complete ACAs.

Figure 6: Assessed quality of Lessons Learned (ACAs)



[19] A12160: Bougainville Elections Assistance
[20] A11698: PF 1-319 Pacific Islands Secondary Tertiary Development Programme
[21] A12513; A12518; A12499; A12512; A12504; A12505; A12515; A125126; A12605: Tropical Cyclone Winston.

Actions to Address Issues (AMAs)

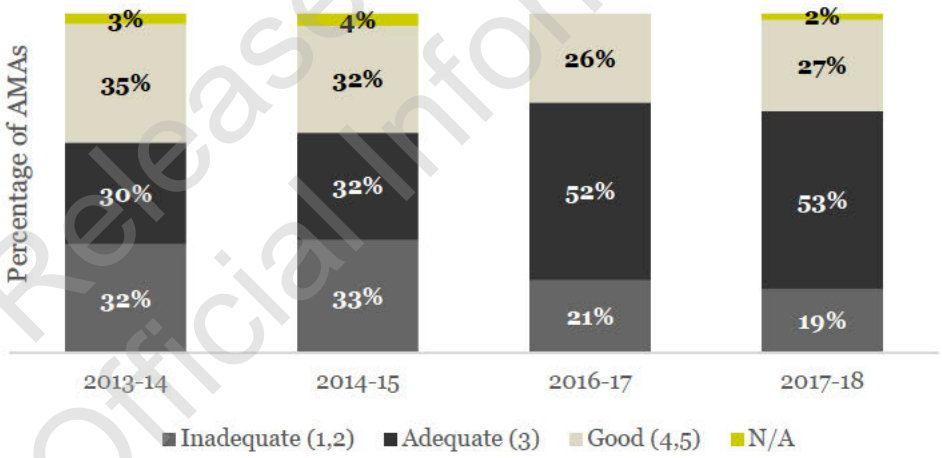
In AMAs, the rating assessing actions to address issues considers the extent to which existing and emerging issues that affect an Activity's progress or performance are consistently identified and followed through to specific management actions that would address them. Ratings of 4 and 5 (good and very good) include analyses that follow existing and emerging issues through with appropriate management actions. Ratings of 2 and 3 apply to vague actions that may not be consistent with emerging and existing issues identified in the AMA.

Compared to the baseline, the proportion of AMAs that identified good or very good actions to address issues has decreased (from 35% to 27%) – see Figure 7, though the result is consistent with the findings in 2016-17. In 16 AMAs, clear and specific management actions were outlined. Additionally, the four Scholarship AMAs identified meaningful actions to address issues, demonstrating insight into technical and management issues that influence these Activities.

There has been an encouraging decrease in the proportion of AMAs that did not identify any actions, or inconsistently identified some actions that may not be substantiated (which, at 19%, is 13% lower compared to the inaugural AAR following a consistent decrease over the previous two AARs). The most common shortcomings were that actions were not consistently identified to address all identified issues; actions to address issues were rarely time-bound and often not sufficiently specific; and actions noted in different sections of AMAs were not consolidated in the appropriate section.

At the same time, AMAs that identified actions to address some of the identified issues increased by 23% compared to the inaugural AAR. Results are consistent with the findings in 2016-17, which were also a significant improvement on the previous year. This suggests that proportionately more AMAs are shifting to 'adequate' usefulness as far as the identification of actions to strengthen Activities is concerned. Further training and technical support could be instrumental to continually enhance the usefulness of AMAs in addressing issues to strengthen the quality of Activities

Figure 7: Assessed quality of Actions to Address Issues (AMAs; Scholarship Activities excluded)



Some AMAs reported progress that was ahead of planning, or targets were being exceeded quite early on. While this would warrant a robust higher rating, it should be noted that few AMAs included actions aimed at understanding the contributing factors to better-than-expected progress, and to amend planning going forward. This could include, for example, considering whether original targets were not sufficiently ambitious and amending planning and RMTs to re-set targets; monitoring implementation to limit unplanned "over-delivery"; ensuring that over-achievement of some targets does not impact negatively on other targets; identifying where reasons for over-delivery constitutes efficiency and where it could be replicated or scaled up in other parts of the Activity, etc.

The AMA for a multi-country Activity missed an opportunity to investigate reasons for varying progress between countries, and how this knowledge could enhance the effectiveness of the Activity across all countries.

Drawing on AMAs and ACAs to improve Activities

The extent to which AMAs and ACAs can be drawn on to improve Activities depends on whether they are based on evidence from more than one source (regardless of the robustness of effectiveness ratings); whether they tell a coherent, evidence-based story of progress; and whether they consistently identify meaningful actions (AMAs) or lessons (ACAs) to inform the improvement and future planning of Activities (rating scales are presented in Appendix 1).

Figure 8 suggests that there could be reason to be more confident in drawing on AMAs and ACAs to improve Activities:

- The percentage of AMAs that received an overall rating of 1 or 2 (meaning they do not contain consistent, evidence-based information to improve the associated Activity) decreased from 43% in the inaugural AAR to 19% in the current AAR; that is a decrease of 24%.

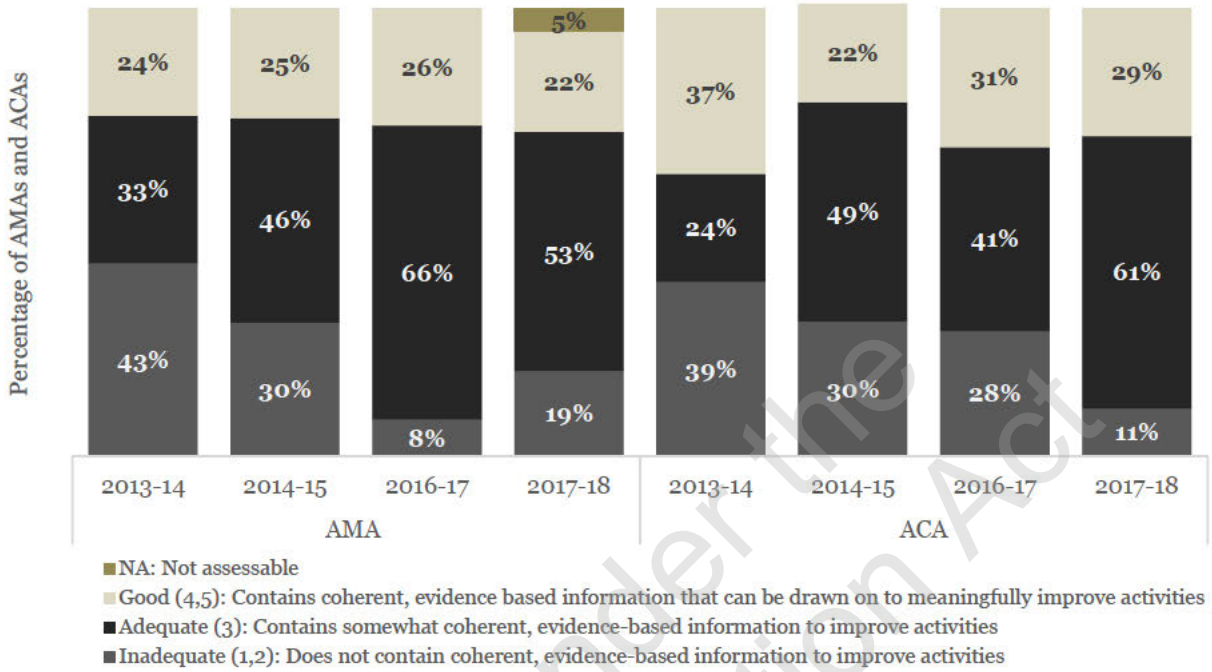
  The percentage of AMAs that received overall ratings of 4 and 5 (meaning that they contain highly consistent, evidence-based information to improve the associated Activity) shows small decreases across successive AARs. In the current AAR, 22% of AMAs received a rating of 4 or 5, which is a decrease of 2% compared to the inaugural AAR. This coincided with an increase of 20% in the proportion of AMAs that contain fairly consistent, evidence-based information to improve the associated Activities increased by 20% - up from 33% in the inaugural AAR to 53% in the current AAR. This indicates that the proportion of AMAs that cannot be drawn on to improve Activities is decreasing in favour of those that can be drawn on to improve Activities, although there is room for improvement.

- Over time, the proportion of ACAs that does not contain consistent, evidence-based information to improve Activities has reduced substantially (from 39% in the inaugural AAR to 11% in the current AAR – therefore a decrease of 28%), but the proportion provides highly consistent, evidence-based information to improve Activities has also decreased somewhat (from 37% to 29% - therefore a decrease of 8%). The proportion that provides fairly consistent, evidence-based information to improve Activities has increased by 37% - up from 24% in the inaugural AAR to 61% in the current AAR. This indicates that the vast majority of ACAs contain some information that could meaningfully be used to improve Activities in the future.

Box 2
2017/18 AMAs on strengthening Scholarship Activities

When assessed separately, all four scholarship AMAs in the current AAR refer to actions that could strengthen monitoring, reporting and AMAs for scholarship Activities in the future. This includes the development of a revised RMT for scholarship activities, which was being undertaken as part of a strategic evaluation of the scholarship programme. The evaluation was underway at the time the AMAs were being drafted. The three country scholarship AMAs also refer to a tracer study that was underway at the time the AMAs were being drafted, as well as the alumni strategy and a new Scholarships and Alumni Management System (SAM), which are expected to strengthen monitoring and reporting against medium-term outcomes. It would be important to monitor the effect of these initiatives on the robustness and quality of scholarship AMAs in the future.

Figure 8: Drawing on AMAs and ACAs to improve Activities

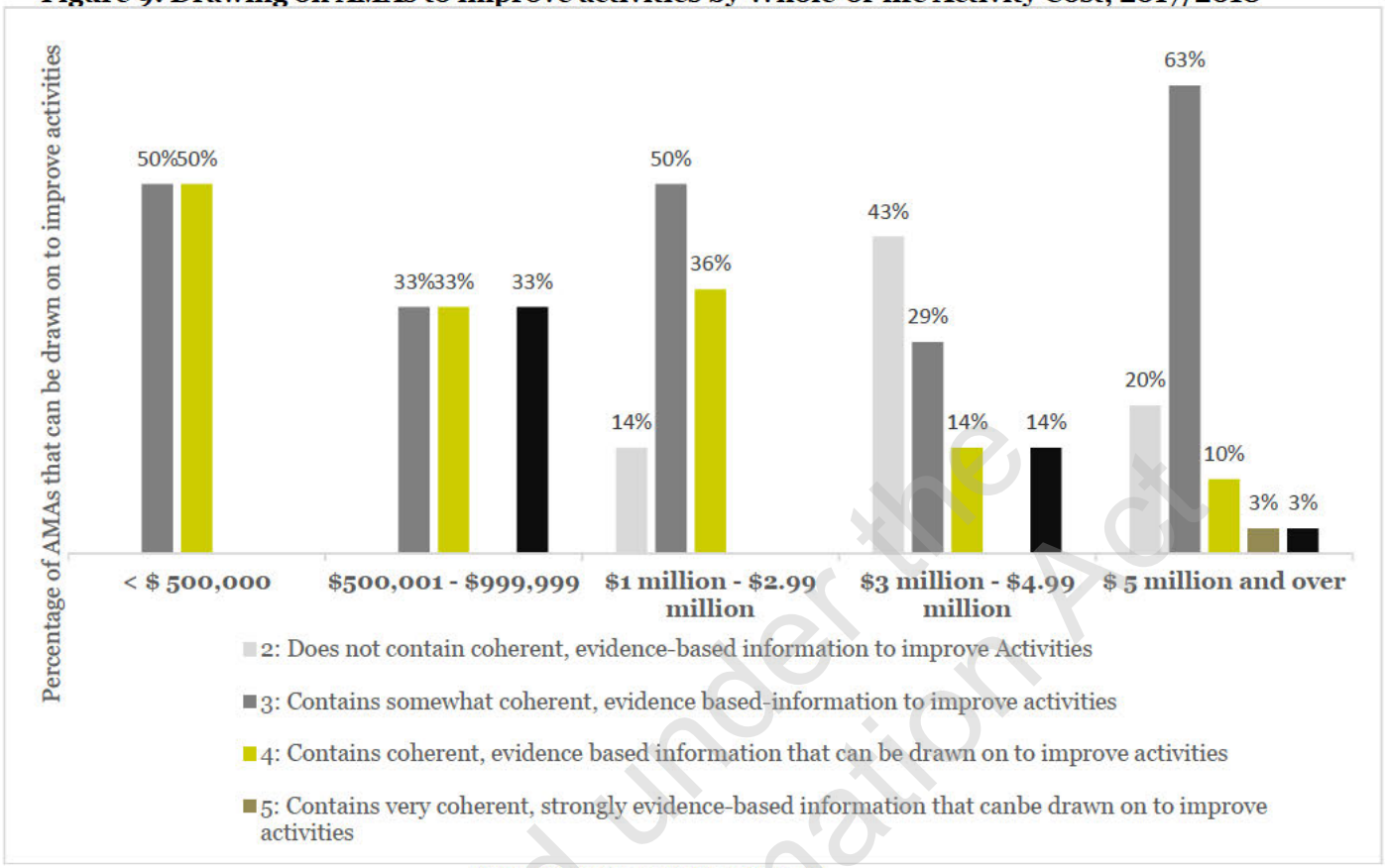


The relationships between the extent to which AMAs/ACAs can be used to improve Activities and the whole-of-life cost (actual) of the corresponding Activities are illustrated in Figures 9 and 10.[22] It can be summarised as follows:

- It appears that the proportion of AMAs that contain consistent, evidence-based information that can be used to meaningfully improve the associated Activities decreases as whole-of-life Activity whole-of-life costs increase. Compared to Activities with smaller whole-of-life costs, fewer AMAs for 'bigger budget' Activities contain consistent, evidence-based information to improve the associated Activities - 43% of AMAs for Activities with whole-of-life costs between $3 million and $4.99 million, and 20% of those with whole-of-life costs of $5 million and over did not contain consistent, evidence-based information to improve the corresponding Activities. This could reflect the challenges associated with the monitoring of complex Activities, as well as funding for Multilateral Organisations, which usually receive substantial funding from MFAT (see Box 1).
- It is encouraging that the only AMA that that received a rating of 5 (very good) for consistently identifying evidence-based issues that affect the progress and performance of the corresponding Activity and following these through with meaningful actions relates to an Activity with a whole-of-life budget of $5 million and over: A12434 - Strengthening Pacific Eye Care System. Across all whole-of-life cost categories, the majority of ACAs contain some consistent, evidence-based information that could improve Activities in the future. The majority of ACAs for Activities with whole-of-life costs between $3 million and $4.99 million (60%) contained highly consistent, evidence-based information that could meaningfully strengthen MFAT's Activities in the future).

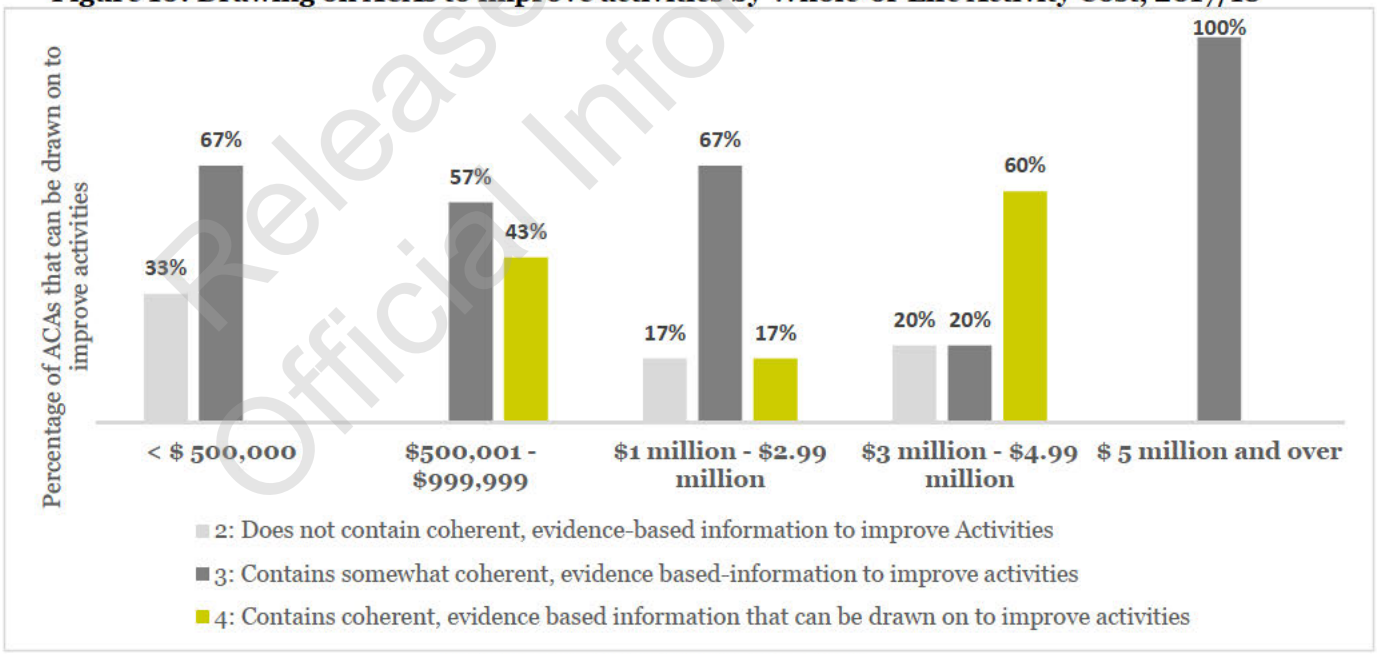
[22] No AMA received and overall quality rating of 1, so this rating has been excluded from Figure 9. No ACA received an overall quality rating of 1, 5 or N/A, so these have been excluded from Figure 10

Figure 9: Drawing on AMAs to improve activities by Whole-of-life Activity Cost, 2017/2018



Note: No AMAs received a rating of 1

Figure 10: Drawing on ACAs to improve activities by Whole-of-Life Activity Cost, 2017/18



Note: No ACAs received ratings of 1, 5 or N/A

4. Summary of Findings

1. Anecdotally, interviews with Activity Managers conducted as part of the current AAR revealed that they have a positive attitude about AMAs and ACAs, as well as a good understanding and appreciation of their value. Fewer Activity Managers regard the completion of AMAs and ACAs as a compliance requirement and there was a clear sense that more Activity Managers view the completion of AMAs and ACAs as important opportunities to reflect analytically on the progress, performance and challenges of Activities and their on-going improvement. This is substantiated by relatively high AMA and ACA completion rates.
2. Using the adjusted post-interview robustness ratings, in AMAs, the robustness of output and medium-term outcomes ratings is lower compared to previous AARs, but the robustness of short-term outcome ratings is higher, reaching the 75% confidence threshold. In ACAs, the robustness of short- and medium-term outcomes is just above the 75% confidence threshold.
3. References to RMTs as sources of evidence for completing AMAs and ACAs are increasing, indicating that Activity Managers are increasingly relying on RMTs to keep track of Activities' progress. The quality of RMTs appears to be improving. However, shortcomings of RMTs continue to hamper monitoring and reporting of results. Many RMTs are not updated and are no longer adequate to monitor progress of evolving Activities. Other major shortcomings of RMTs include the absence of baselines, targets and data to monitor and report progress.

   AMAs of funding for Multilateral Organisations appear to be especially challenging because the results frameworks and progress reporting of these Organisations do not always correspond to MFAT's requirements. It appears that developing AMAs and ACAs for complex Activities, for example multi-county and multi-donor Activities, as well as Activities funded through Budget support could also be challenging because RMTs for these Activities are not straight-forward.

   It is encouraging that an increasing number of Activity Managers are proposing appropriate actions to address the identified shortcomings of RMTs.
4. The identification and analysis of Cross-Cutting Issues remain a challenge for Activity Managers. Although the quality of analysis of CCIs in both AMAs and ACAs is showing improvement, the proportion of these analyses that received 'good' or 'very good' ratings remains small. There was also an increase in the proportion of AMAs and ACAs where CCI analyses received N/A ratings, suggesting that there either were no cross-cutting priorities identified in the AMA or ACA (i.e. CCI is classified as 'not measured' or 'not targeted'), or there was very limited or no further elaboration on CCIs in the document..
5. Observed improvements in the quality of more analytical aspects of AMAs and ACAs are encouraging. Compared to the baseline, the proportion of ACAs that identified well-substantiated, useful lessons increased by 19%, while proportionately more AMAs are shifting from 'inadequate' to 'adequate' quality as far as the identification of actions to strengthen Activities is concerned.
6. There is reason to be cautiously confident in the overall usefulness of AMAs and ACAs. The majority of AMAs and ACAs are of adequate usefulness. The proportions of AMAs and ACAs that are of inadequate usefulness have decreases substantially compared to the baseline – by 24% in the case of AMAs and by 20% in the case of ACAs. These appear to have shifted mainly from 'inadequate' to 'adequate' usefulness. However, a small proportion of AMAs and ACAs appear to be slipping back from 'good' to 'adequate' usefulness.

5. Conclusions

1. Despite improvements in qualitative aspects of reporting, AMAs and ACAs still do not provide sufficiently accurate, complete or stand-alone records of the activity. As in previous AARs, Activity Managers draw on evidence from a range of sources to assess the effectives of Activities but tend not to document all this evidence in AMAs and ACAs. If evidence is not comprehensively

documented, the loss of institutional knowledge leaves substantial gaps, especially where staff turnover is high. When these gaps build up year-on-year, new Activity Managers might find it challenging to complete insightful AMAs and ACAs, thereby jeopardising the robustness of AMAs and ACAs in the longer term.

2. Despite remaining challenges around the robustness of effectiveness ratings, AMAs and ACAs generally include some consistent, evidence-based information that could be drawn on to improve Activities. Providing useful information to improve complex Activities, which often have high whole-of-life costs, are challenging since ways that could improve these Activities may not be within MFAT's full control.
3. So far, AARs have not revealed major statistically significant results related to AMA and ACA improvement over time. Some trends are beginning to emerge, while valuable lessons have contributed to a much-refined methodology. While MFAT does not expect linear improvement due to contextual factors such as organisational capacity and incentives, over a longer time period statistically significant changes may become evident.

6. Recommendations

1. AMAs and ACAs should remain as essential building blocks of the Aid Programme's performance management system. Activity Managers use AMAs and ACAs to reflect and assess the progress, performance and challenges of Activities and they serve as important repositories of institutional memory and continuity during Activity implementation. Increasingly insightful and usable lessons and actions to address issues, if harnessed through a robust knowledge management system, could also prove valuable in strengthening Activities.
2. Ongoing training and technical support would be important to ensure that gains made in the robustness and usefulness of AMAs and ACAs are maintained and enhanced. Gradual improvements are becoming evident, but it would be important to address known challenges and strengthen capacity to maintain this positive momentum and to prevent the gains made from being lost.

   Continue to provide training and guidance for Activity Managers in Wellington and at Post (including locally-engaged staff) to ensure that they understand why and how to document the evidence base for AMAs and ACAs fully, yet concisely, to increase the proportion of AMAs and ACAs that provide stand-alone records of Activities' progress and performance. This would be instrumental to lift the robustness and usefulness of AMAs and ACAs (and therefore their value as essential building blocks of the Aid Programmes performance management system) to a higher level.

   Training and support in the following priority areas could be considered:

   - Documenting consolidated evidence from several sources to justify effectiveness and DAC criteria ratings.
   - RMT quality and wider socialisation of RMTs as foundations of Activity design, monitoring and reporting, as well as dynamic tools for Activity improvement.
   - MERL expert assistance to support regular reviewing and updating of RMTs to ensure that they remain relevant and up-to-date.
   - Strengthening RMTs and monitoring of complicated and complex Activities, for example multi-donor and multi-country Activities, as well as Activities funded through budget support.
   - Improving consistency and coherence in AMAs and ACAs, including identifying issues that affect the progress and performance of Activities, and following this through into meaningful, evidence-based issues to improve on-going activities (in AMAs), or lessons relevant to comparable types of Activities when they complete ACAs.

3. Given the size of MFAT's funding to Multilateral Organisations and the unique arrangements around their monitoring, it could be beneficial to tailor guidance for the AMAs of these Activities.
4. Provide support to Activity Managers to identify and perceptively address appropriate Activity cross-cutting markers:
    - Where a cross-cutting marker is identified as relevant, it should be dealt with consistently and perceptively throughout the design, monitoring and reporting of the activity, including in AMAs and ACAs.
    - Avoid including cross-cutting markers that are not relevant to an Activity in its AMA/ACA.
5. AARs should continue to be conducted on a periodic basis to monitor the effect of known enablers and constraints to the robustness and usefulness of AMAs and ACAs, as well as to identify emerging challenges and actions for their continuous improvement. A larger database will also enable meaningful trend analyses of the robustness and usefulness of AMAs/ACAs across different sectors, programmes and budget levels.

Appendix 1: Methodology

Sampling

The AAR was conducted on a statistically representative sample of AMAs and ACAs that met the following criteria:

- AMA and ACA submitted between July 2017 and July 2018
- Assessments were conducted within 12 months after the assessment period ended

A total population of 141 AMAs and 55 ACAs met these criteria. A simple random sample of 66 AMAs and 36 ACAs (95% confidence levels, 10% confidence intervals). The sample was selected by an IOD PARC statistician, in consultation with MFAT. Ultimately, the AMA sample was reduced to 63 (that is 45% of AMAs) due to unavailability of information. The ACA sample was also reduced to 28 (51% of ACAs) for the same reason. Statistical calculations were adjusted accordingly. To ensure consistency with previous AARs, AMAs for Scholarship Activities were excluded from all comparative analyses, meaning that comparative analyses are based on an AMA sample of 59.[23]

As with 2016-17, MFAT made technical support available to Activity Managers to support development of AMAs and ACAs. It was important that the review team remained 'blind' as to which Activity Managers had received this support and those which had not. Therefore, the AMA and ACA samples were provided to the review team without this data. After the reviews were completed, MFAT provided this information to the review team to enable a comparative analysis to determine if technical support influenced the robustness and quality of AMAs and ACAs. However, only seven AMAs and one ACA was found to have received technical support. Therefore, an analysis of the influence of technical support on AMA and ACA robustness and quality was not meaningful.

Ratings Approach

The rating approach for assessing **effectiveness ratings** in AMAs and ACAs, as well as ratings for **other DAC criteria** in ACAs:

Rating assessed as robust	Rating assessed as non-robust
A rating was assessed as robust if the reviewer agreed with the assessment of the author, based on the consideration of available evidence (e.g. author rating of 3 and reviewer rating of 3 = R) Reviewers agreed with the authors of Where a N/A (no rating) was assessed as robust, the absence of an RMT prevented a rating from being substantiated and verified, or there was insufficient evidence available to support justification of a rating.	A non-robust rating was given if the reviewer did not agree with the assessment of the author, based on the consideration of available evidence (e.g. author rating of 3 and reviewer rating of 4 = L or author rating of 3 and review rating of 2 =H). Where a N/A (no rating) was assessed as non-robust, the author of an AMA/ACA gave an N/A for reasons other than the absence of an RMT, but based on available evidence the reviewers believed a rating could have been given.

[23] The Scholarships Programme completed AMAs for the first time in 2014/15. Four Scholarship Activities were included in the 2016 AAR and eight were included in the previous AAR. The four scholarship activities included in the current AAR sample account for 3% of the AMA sample.

Ratings Scales for Effectiveness and other DAC Criteria

Rating scale for **Effectiveness Assessments** (AMA and ACA) and **Relevance, Efficiency, Impact and Sustainability** (ACA)

In accordance with the 5-point rating scale in MFAT's Activity Quality Policy[24], effectiveness (AMA and ACA) and relevance, efficiency, impact and sustainability (ACA) were given a rating of between 1 and 5.

Rating	Measure
1	Poor (needs to be profoundly changed)
2	Not adequate (needs to be profoundly improved)
3	Adequate (needs some identified work to improve)
4	Good (needs some minor work in some areas to improve)
5	Very Good (needs ongoing management and monitoring only)
N/A	Insufficient information is available to assess the quality of an AMA according to these measures

Ratings Scales for other AAR quality Criteria

Where the Activity Manager was not required to provide a rating, but simply provided a narrative analysis, the reviewers constructed rubrics to assess the quality of the reporting. This applies to:

- Cross-cutting issues (AMAs and ACAs)
- Actions to address issues (AMAs and ACAs)
- Lessons learned (ACAs only)
- Activity Results Framework (AMAs only)
- Overall usefulness ratings (AMAs and ACAs)

The ratings scales/rubrics for assessing these components of AMAs and ACAs are as follows:

[24] Activity Quality Policy ID: RFE-21-134, last updated February 2015

1. Rating Scale for assessing **Cross Cutting Priorities** (AMA and ACA)

The 2018 AAR Cross-Cutting Priorities rating scale and measures are largely the same as those used for the 2017 AAR, aside from some minor updates to wording for Measures 1, 2, 3 and 4. These have no material change to the assessment method.

Criterion	Rating	Measure
Cross Cutting Priorities	1	Assessments provide no evidence or analysis of how relevant cross-cutting issues and associated risks are dealt with
	2	The appraisal of cross-cutting issues and risks does not draw on any evidence and lacks any depth/ meaningful insight and/or does not correspond with the available evidence
	3	The appraisal of cross-cutting issues and their related risks draws on some evidence (may only be one source) and identifies at least some cross-cutting issues that need to be managed/are being managed.
	4	The appraisal of cross-cutting issues draws on sound evidence base and provides insight into the most critical dimensions of cross-cutting issues that need to be managed/are being managed.
	5	The analysis of cross-cutting issues draws on multiple, triangulated sources of evidence, and consistently provides the reader with meaningful insights into how all identified cross-cutting issues are currently being managed/not managed
	N/A	Not assessable based on the evidence provided.

2. Rating scale for assessing **Actions to Address Issues** (AMAs)

The 2018 AAR Actions to Address Issues rating scale and measures are largely the same as those used for the 2017 AAR, aside from some minor clarity updates to wording for all measures. These have no material change to the assessment method.

Criterion	Rating	Measure
Actions to address issues (Rating scale for AMAs)	1	Does not identify any future oriented actions although there are clear performance issues in the report.
	2	Identifies some future oriented actions, but these are not specific, clear and do not appear to directly address the issues in the report.
	3	Identifies existing and emerging issues in some sections and somewhat addresses these throughout the report.
	4	Identifies existing and emerging issues in most sections and addresses these with clear and specific management actions.
	5	Consistently identifies existing and emerging issues (across relevance, effectiveness, efficiency, cross cutting issues etc) that require management attention and addresses these coherently and comprehensively with specific management actions (owned and timebound).
	N/A	Not assessable based on evidence provided

3. Rating scale for assessing **Lessons Learned** (ACAs)

The 2018 AAR Lessons Learned rating scale and measures are largely the same as those used for the 2017 AAR, aside from some minor clarity updates to wording for measures 2, 3, 4 and 5. These have no material change to the assessment method.

Criterion	Rating	Measure
Lessons learned (Rating scale for ACAs)	1	Does not identify any lessons learned
	2	Identifies generic lessons that don't obviously correlate with program findings
	3	Identifies some generic lessons learned that could inform future programming, that logically flow from program findings
	4	Identifies a number of specific lessons learned that could be used to inform future planning
	5	Consistently identifies meaningful and specific lessons learned that could inform future programming and that demonstrate thoughtful insight and reflection.
	N/A	Not assessable based on evidence provided

4. Rating scale for the **extent to which AMAs and ACAs can be drawn on to improve Activities** [25]

In accordance with the 5-point rating scale in MFAT's Activity Quality Policy, each AMA and ACA was given a rating of between 1 and 5 to signify whether it contains consistent, evidence-based information that can be drawn on to improve Activities. The 2018 AAR Lessons Learned rating scale and measures are largely the same as those used for the 2017 AAR, aside from some minor clarity updates to wording for all measures. These have no material change to the assessment method.

Criterion	Rating	Measure
Report Quality	1	The AMA/ACA lacks an evidence base and coherence. It does not identify action/lessons to inform the improvement and future planning of activities.
	2	Most ratings and appraisals lack evidence, or only draw on one source of evidence. Overall, the AMA/ACA lacks coherence. Actions/lessons learned are not meaningful to inform improvement and future planning of activities.
	3	Some ratings and assessments are based on evidence from more than one source. The AMA/ACA is fairly coherent. Some of the actions/lessons learned are meaningful to inform improvement and future planning of activities.
	4	Most ratings and assessments are based on evidence from multiple sources. The AMA/ACA is coherent and most of the actions/lessons learned are meaningful to inform improvement and future planning of activities.
	5	All ratings and assessments are based on triangulated data from multiple sources. The AMA/ACA tells a coherent story of the activity's performance. It identifies highly meaningful actions/lessons to inform the improvement and future planning of activities.
	N/A	Not assessable based on evidence provided

[25] In previous AARs, 'this was referred to as 'Report Quality'

5. Rating scale for assessing **Activity Results Frameworks** (AMA)

The 2018 AAR Lessons Learned rating scale and measures are somewhat changed from those used for the 2017 AAR. There have been some minor clarity updates to wording for measures 2, 3, 4 and 5. The wording of 2018 measure 5 clarified that the rating is only applied in exceptional, leading practice circumstances, which is in accordance with the intent in 2017. There is a change in focus used for the assessment method for ratings 3 and 4. In 2018 measure 3, the Activity Manager will make appropriate recommendations to address a flawed RMT, whereas in 2017 measure 3 they did not make appropriate recommendations to address a flawed RMT.

Criterion	Rating	Measure
Quality of activity's Results Framework / logic and reporting and how this affects ability to track and substantiate progress and performance (Rating scale for AMAs)	1	Activity does not have Results Framework / Logic. The Activity Manager does not make appropriate recommendations to address this
	2	Results Framework / Logic has shortcomings which affect the quality of the AMA. The shortcomings are not acknowledged by the Activity Manager and appropriate recommendations are not made to address this
	3	Activity does not have Results Framework, or the Results Framework/Logic has shortcomings which affect the quality of the AMA. The implementing partner may not use it for reporting and/or the activity manager may not draw on evidence from partner reports to assess the activity's progress and performance. The Activity Manager makes appropriate recommendations to address this.
	4	The Activity has an adequate Results Framework/Logic. The Activity Manager can draw on evidence from partner reports to assess the Activity's progress and performance, based on the Results Framework/Logic.
	5	The Activity Manager draws on a well-developed Results Framework/Logic and consistent partner reporting to make judgements about the activity's progress and performance. Results framework can be used as a good practice example.
	N/A	Cannot assess extent to which Results Framework / Logic and progress reporting affect quality of AMA.

Interviews

Where an effectiveness rating was given by the Activity Manager, the assessment team determined whether this rating appeared robust, high or low based on evidence provided in the AMA/ACA, in relation to the Activity Quality Policy standard ratings.

Interviews were indicated by the following:

- Where an effectiveness rating was assessed as non-robust, and
- Where a results framework was available to make a valid judgement against.

If, on interview, further evidence was provided to justify the Activity Manager's rating, the AAR rating was adjusted to robust. If no further robust evidence is provided, the original rating of high or low remains. Further evidence may have been presented which would change the rating, but still mean that overall the rating of a high or low remained.

Qualitative Analysis

Qualitative analysis of findings was conducted through the following approach:

- From interviews, the assessors documented Activity Managers' justification of effectiveness ratings that were assessed as non-robust in the evidence spreadsheet. A content analysis of reasons was conducted to identify common themes/reasons on the

basis of which the final rating was decided. Each reason was then coded to a common theme/reason and a frequency analysis of theses/reasons was conducted.

- Information on Activity's Managers' process when completing AMAs/ACAs, their opinion of the AMA/ACA templates and guidance, their view of the robustness of evidence available to inform the completion of AMAs and ACAs, as well as which aspects of AMAs/ACAs they found especially challenging (and why) was documented. A content analysis of the documented information was conducted to identify commonalities and emerging themes.

Sample Revision for AAR Quantitative Analysis

Quantitative Analysis

Comparative assessments of results between the four AARs was completed. Robustness of effectiveness ratings at output, short term outcome, medium term outcome for AMAs and ACAs as well as relevance, efficiency, impact and sustainability for ACAs was compared and presented in charts. For the remaining criteria (Cross-cutting issues (AMAs and ACAs) , Overall quality ratings (AMAs and ACAs), Actions to address issues (AMAs only), Activity Results Framework (AMAs only), and Lessons learned (ACAs only)), the assessed ratings were compared by percentage and presented in charts for inadequate (rating of 1 or 2), adequate (rating of 3) and good (rating of 4 or 5).

N/A Ratings

N/A ratings were removed from the observed proportion to enable the calculation of robustness.

Scholarships Programme

Scholarships Activities were introduced to the AAR process in 2016 (AMAs for 2014-15) and have a notable influence on the overall robustness of AMA effectiveness ratings. For purposes of comparability across all four AARs, assessments of the robustness of AMAs in this report consistently exclude Scholarship Activities unless noted otherwise.

Statistical Analysis of results

The Wilson score method was used to construct initial confidence intervals[26]:

$$\frac{2np + z^2 \pm z\sqrt{4npq + z^2}}{2(n + z^2)}$$

In this calculation n is the sample size, p is the observed proportion, q=1-p and z is the quintile of the standard normal distribution which depends on the desired level of confidence. For 95% confidence intervals, as is the case in this instance, z is 1.96. Using the same formulas as in previous AARs[27], these were the adjusted using the 'finite population correction' method referenced by Burstein[28] to construct confidence intervals with more accurate coverage probability given the small population sizes. This approach enables the nominal coverage of 95% to be maintained while narrowing the confidence intervals.

To determine **statistical significance** of differences in findings between the inaugural and 2017-18 AAR and the 2016-17 and 2017-18 AAR, the Z test based on normal approximation in two finite populations was used[29]. It was possible to conduct statistical analyses when a rating was applied by

[26] Newcombe, R.G. (1998) Two-sided confidence intervals for the single proportion: comparison of seven methods. Stat. Med. 17, 857-872.

[27] Wilson's method alone was used for the inaugural 2015 AAR report (2013-14 results). The 2016 AAR report (2014-15 results) used the current combination of Wilson overlaid with the FPC method referenced in Burstein; the 2013-14 results were revised accordingly for comparison. See 2016 AAR Report for further detail.

[28] Burstein. H (1975), Finite Population Correction for Binomial Confidence Limits, Journal of the American Statistical Association, Vol. 70, No. 349 (Mar., 1975), pp. 67- 69; Rosenblum, M, van der Lann. M.J (2009) Confidence Intervals for the Population Mean Tailored to Small Sample Sizes, with Applications to Survey Sampling, Int. J Biostat 5(1):4

[29] Krishnamoorthy K and Thomson J. Hypothesis Testing About Proportions of Two Finite Populations. The American Statistician, Vol 56, No 3 (August 2002), pp 215-222.

the Activity Manager. This applies to:

- Effectiveness ratings at output, short term outcome, medium term outcome for AMAs
- Relevance, effectiveness ratings (output and short- and medium-term outcomes), efficiency, impact and sustainability for ACAs.

Additional Analysis in 2018

Additional analyses that were carried out in the current AAR that were not included in previous AAR are the following:

- Comparison of 2018 AAR results against Business Units and Divisions to identify if:
  - there are any significant differences in the **overall quality** of AMAs and ACAs between Business Units (Program)?
  - There are any significant differences in the **overall quality** of AMAs and ACAs between Divisions (Sector/Investment Priority)?

Adjusted Effectiveness Ratings

To understand the possible effect of the low interview rates in the 207-18 AAR, adjusted post-interview robustness ratings were applied to the initial post-interview AAR results. The adjustment for each result area is based on the average percentage change in post-interview robustness of effectiveness ratings across the previous three AARs and applying it to the pre-interview rating in the current AAR.

Due to the comparatively low interview rate compared to previous AARs, adjusted robustness rates for effectiveness ratings were calculated, based on the average percentage change in the robustness of effectiveness ratings across three previous AARs, as illustrated in Table A-1 below.

Table A-1: Percentage Change between Desk-based and Post Interview Effectiveness Robustness Ratings

Key Criteria	AMAs			ACAs	
	Outputs	Short Term Outcomes	Medium Term Outcomes	Outputs	Short- and Medium-Term Outcomes
Inaugural AAR	38%	35%	9%	34%	23%
2014/15 AAR	39%	23%	15%	33%	8%
2016/17 AAR	59%	68%	50%	35%	29%
Total increase across three AARs	136%	126%	73%	103%	60%
Average increase across three AARs *multiplier applied to AAR 2017-18 results	45%	42%	24%	34%	20%
Increase in Pre-Adjustment AAR 2017-18	15%	10%	9%	13%	11%
Application of multiplier					
AAR 2017-18 pre-interview Robustness Ratings	44%	53%	59%	54%	64%
Adjusted Robustness Ratings	64%	75%	74%	72%	77%

For AMAs, these average increases, of 45%, 42% and 24% for output, short-term outcome and medium-term outcome AMA ratings respectively, were then applied as a multiplier to the pre-interview robustness of effectiveness ratings in the current AARs.

For ACAs, interview rates of 86%, 92% and 84% resulted in average changes of 34% and 20% in the robustness of output and short- and medium-term effectiveness ratings, respectively.

The difference between the desk-based ratings and post-interview effectiveness robustness ratings (unadjusted) are included in Table 2 for AMAs and Table 3 for ACAs.

Table A-2: Difference between Desk-based and Post Interview Effectiveness Robustness Ratings: AMAs, unadjusted

Effectiveness Criteria	AAR	Pre-Interview	Post Interview ratings	Difference
Outputs	2013-14	50%	69%	19%
	2014-15	60%	84%	24%
	2016-17	48%	75%	27%
	2017-18	44%	51%	7%
Short-term outcomes	2013-14	53%	72%	19%
	2014-15	49%	60%	11%
	2016-17	41%	69%	28%
	2017-18	53%	58%	5%
Medium-term outcomes	2013-14	58%	63%	5%
	2014-15	61%	70%	9%
	2016-17	49%	74%	25%
	2017-18	59%	64%	5%

Table A-3: Difference between Desk-based and Post Interview Effectiveness Robustness Ratings: ACAs, unadjusted

Effectiveness Criteria	AAR	Pre-Interview	Post Interview ratings	Difference
Outputs	2013-14	69%	93%	24%
	2014-15	67%	89%	22%
	2016-17	53%	72%	19%
	2017-18	54%	61%	7%
Short-and Medium-term outcomes	2013-14	69%	85%	16%
	2014-15	69%	75%	6%
	2016-17	66%	84%	18%
	2017-18	64%	71%	7%

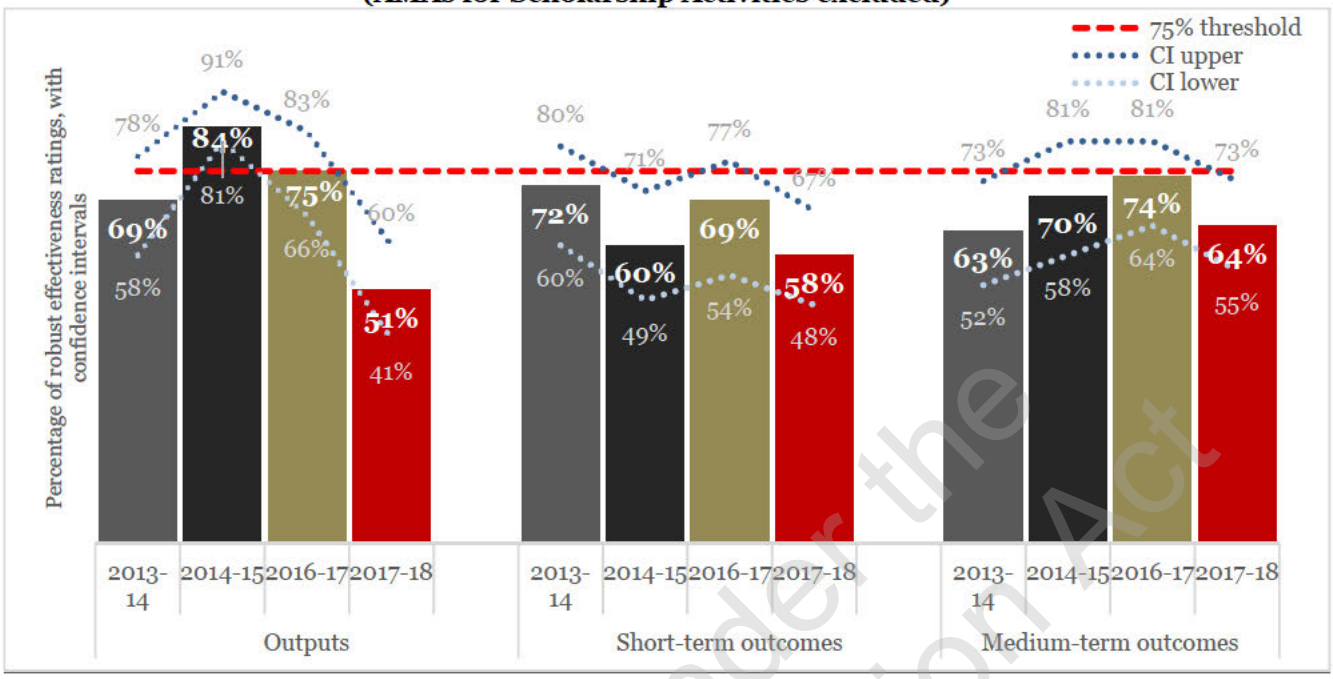
Initial Effectiveness Findings, based on pre-adjusted Effectiveness Ratings

For completeness, the initial, pre-adjusted effectiveness findings are described in this section.

Figure A-1 describes the robustness of post-interview effectiveness ratings for outputs and outcomes in **AMAs** (excluding scholarship AMAs). None of the post-interview effectiveness ratings meets MFAT's confidence threshold of 75%. The robustness of all effectiveness ratings has decreased compared to the 2016/17 AAR, and that of outputs and short-term outcomes is also lower compared to the inaugural AAR. The decrease in the robustness of output ratings compared to the previous two AARs is statistically significant. Across the board, confidence intervals (and margins of error) are slightly smaller compared to the inaugural AAR.[30]

[30] The percentage of robust output ratings decreased from 69% in the inaugural AAR to 51% in the current AAR, while the confidence intervals - and therefore the margin of error - decreased slightly from 20% to 19%. The percentage of robust short-term outcome ratings decreased from 72% in the inaugural AAR to 58% in the current AAR, while the confidence intervals – and therefore the margin of error – decreased slightly from 20% to 19%. The percentage of robust medium-term outcome ratings increased from 63% in the inaugural AAR to 64% in the current AAR, while the confidence intervals – and therefore the margin of error – decreased from 21% to 18%.

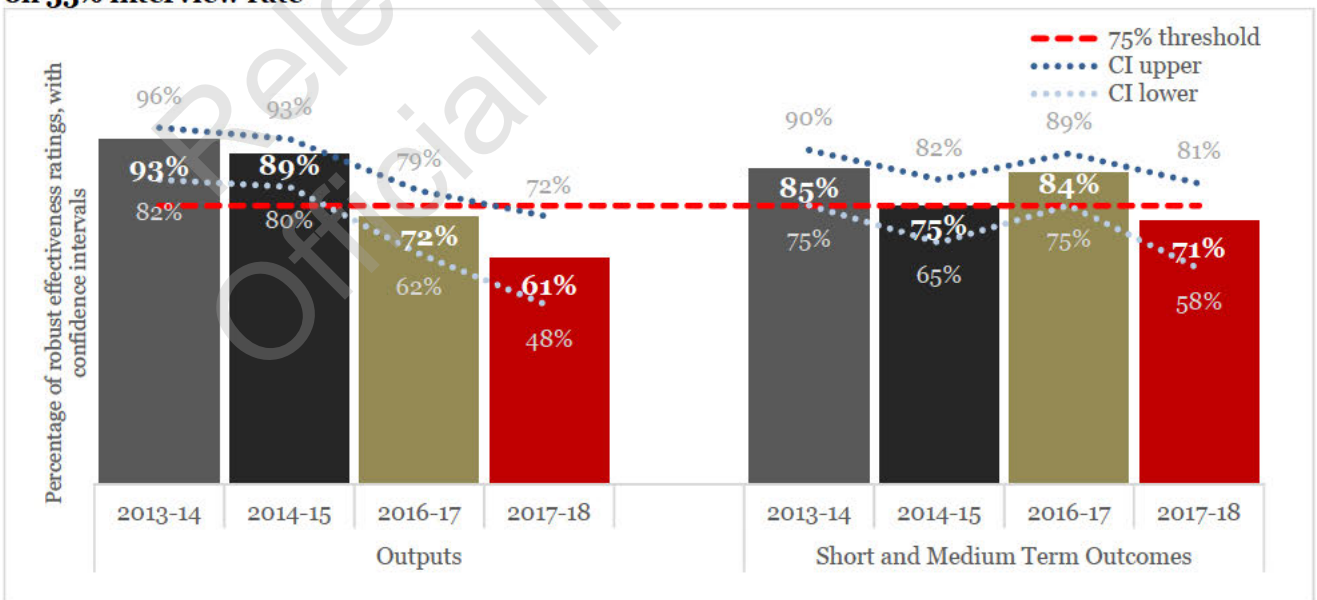
**Figure A-1: Robustness of post-interview effectiveness ratings in AMAs across four AARs, based on 33% interview rate*
(AMAs for Scholarship Activities excluded)**



* These figures are based on a relatively low interview rate of 33%, compared with interview rates of 86% in the inaugural AAR; and 92% and 84% in the two subsequent AARs.

As indicated in Figure A-2, for the second consecutive time, there has been a statistically significant decrease in the robustness of post-interview output ratings in ACAs compared to the inaugural AAR. For the first time, the robustness of short- and medium-term outcome ratings in ACAs is below the 75% confidence threshold. However, in both cases the margin of error is higher compared to previous AARs.[31]

Figure A-2: Robustness of post-interview effectiveness ratings in ACAs across four AARs, based on 33% interview rate



[31] For outputs, the margin of error is 24, compared to 14 in the inaugural AAR; 13 in the second and 17 in the previous AARs. For short- and medium-term outcomes, the margin of error is 23, compared to 15 in the inaugural AAR; 17 in the second and 15 in the previous AAR.

Appendix 2: Terms of Reference

Description of Services

The New Zealand Aid Programme uses a results-based approach for designing and managing Activities. For each New Zealand Aid Programme Activity, an Activity Results Framework (ARF) is used to track progress towards intended results.

Activity Monitoring Assessments (AMAs) are completed annually through the life of an Activity, and Activity Completion Assessments (ACAs) are completed after completion of an Activity. These assessments are key internal mechanisms for assessing results and lessons learned against Activity Results Frameworks. AMAs and ACAs are written by MFAT Activity Managers and draw on evidence from reports from implementing partners, supplemented by any monitoring visits, and feedback from other MFAT staff (e.g. Post) and stakeholders. These reports include results data, related commentary and ratings against specified criteria. AMAs and ACAs are an opportunity for Activity Managers to critically reflect on the progress and achievements of their Activity. They are the building blocks of the New Zealand Aid Programme performance system, and are crucial to enabling results-based management of Activities by Programme teams.

Description of Services

Annual Assessment of Results

- The "Annual Assessment of Results" (AAR) is a quality assurance mechanism via an independent check of New Zealand Aid Programme results and ratings. The inaugural AAR was conducted in early 2015 on a representative sample of AMAs and ACAs from the 2013/2014 financial year. This provided a baseline for future assessments and confirmed the AAR methodology. A second AAR was conducted in 2016 on a sample of AMAs and ACAs from 2014/15 financial year, while a third AAR was conducted in 2018 on a sample of AMAs and ACAs from the 2016/17 financial year.

- The same methodology for sampling as previous years will be used this year, with an extended sample for an additional review of AMAs and ACAs completed by Activity Managers who received training or advisory support from MFAT's Planning and Results team. Approximately the same total number of AMAs and ACAs will be reviewed as previous years.

- This AAR will contain the following specific analyses:

  (1) Comparison of the findings with those of the previous three AARs;

  (2) Comparison of results with Scholarships included and excluded;

  (3) Comparison of AMAs and ACAs that were completed using the old templates and those that were completed using the revised templates;

  (4) Comparison of AMAs and ACAs that received technical support with those that did not.

  Items 3 and 4 are additional analyses not undertaken in the previous two years.

- Deliverables covered by this contract include:
  - Desk based review of selected AMAs and ACAs
  - Completion of interviews with a sample of Activity Managers
  - One workshop with MFAT on emerging findings and report structure
  - One short report detailing the assessment findings and recommendations delivered in a professional manner and in accordance with MFATs internal 'Style Guide'

Milestones

Annual Assessment of Results		
No.	Output/milestone	Timeline
1	Finalise inception and design stage including • Finalise sampling approach • Update and finalise AMA and ACA review templates in consultation with MFAT • Draw sample • Review 2 AMAs and 1 ACA to test template and calibrate interpretation of questions and ratings • Discussion to standardise and make final adjustments to templates • Document methodology including sampling approach	10 December 2018
2	Desk based review of remaining AMAs and ACAs • Including quality assurance of a selection of assessments	7 January - 11 February 2019
3	Interviews of sample of Activity Managers • Finalise script and approach for telephone interviews • Conduct telephone interviews to clarify non-robust ratings • Interview write-up and analysis. Finalise ratings accordingly.	4 – 25 March 2019
4	Workshop with MFAT on emerging findings and report structure	17 April 2019
5	Draft report produced, and report finalised • Synthesis of findings from analysis, interviews and workshop • Incorporate MFAT feedback on draft report	17 May 2019

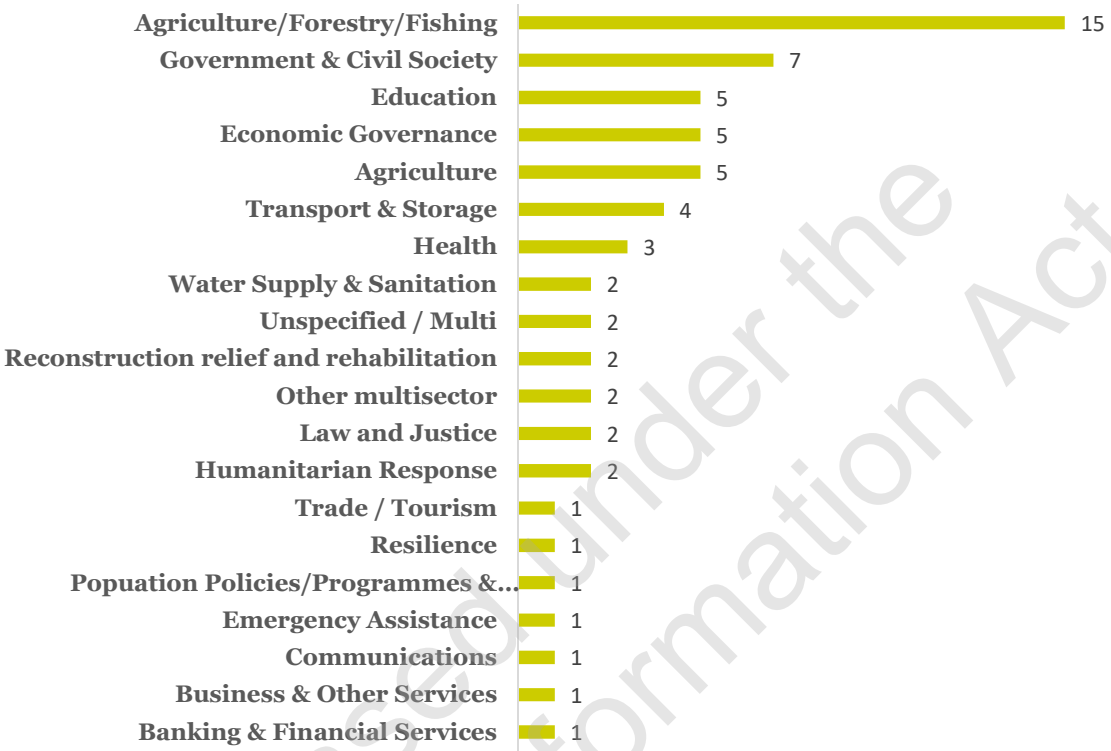
Performance standards

- The reports produced are in line with the Ministry's 'Style Guide'. This will be included as Appendix One.
- The reports produced are aligned with the content of the Terms of Reference document provided with the Request for proposal. The Terms of Reference will be included as Appendix Two. The reports produced are accepted by the Buyer.
- The reports produced will be crafted with due care, skill and diligence, and to the appropriate professional standard or in accordance with good industry practice as would be expected from a leading supplier in this industry.

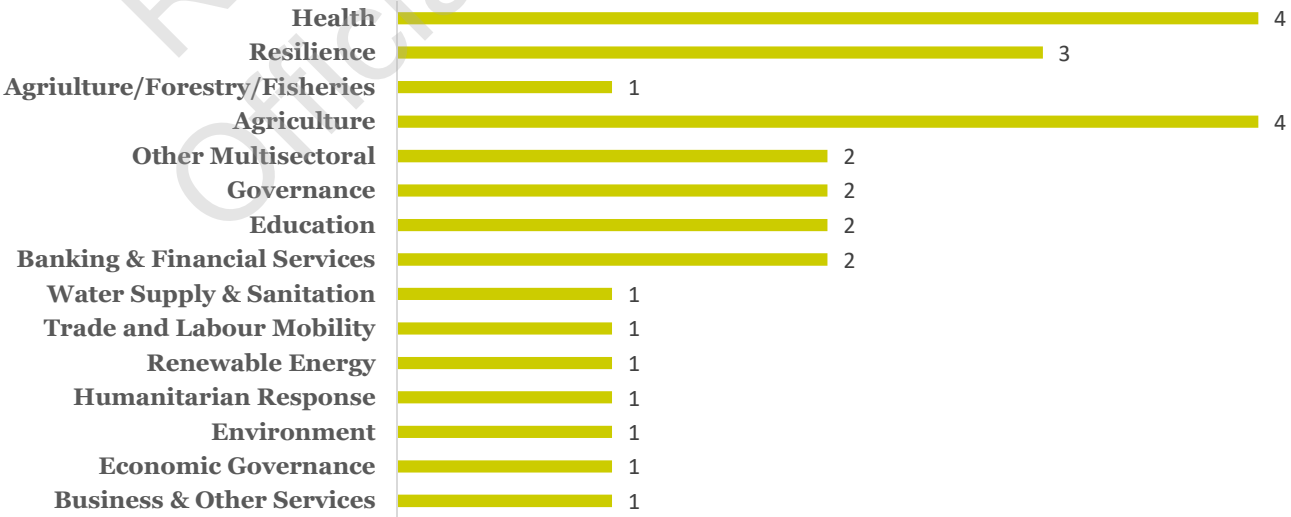
Appendix 3: Distribution of sample

The following figures profile some of the key descriptors for the AMA and ACA samples:

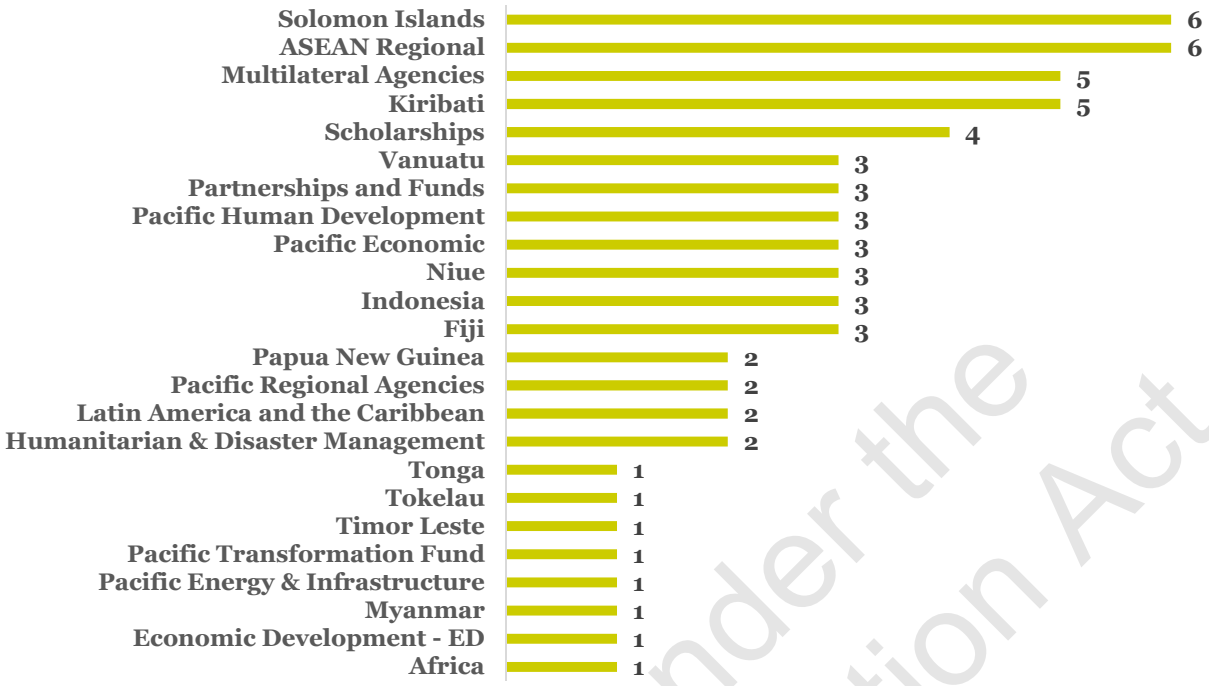
Number of AMAs by Sector



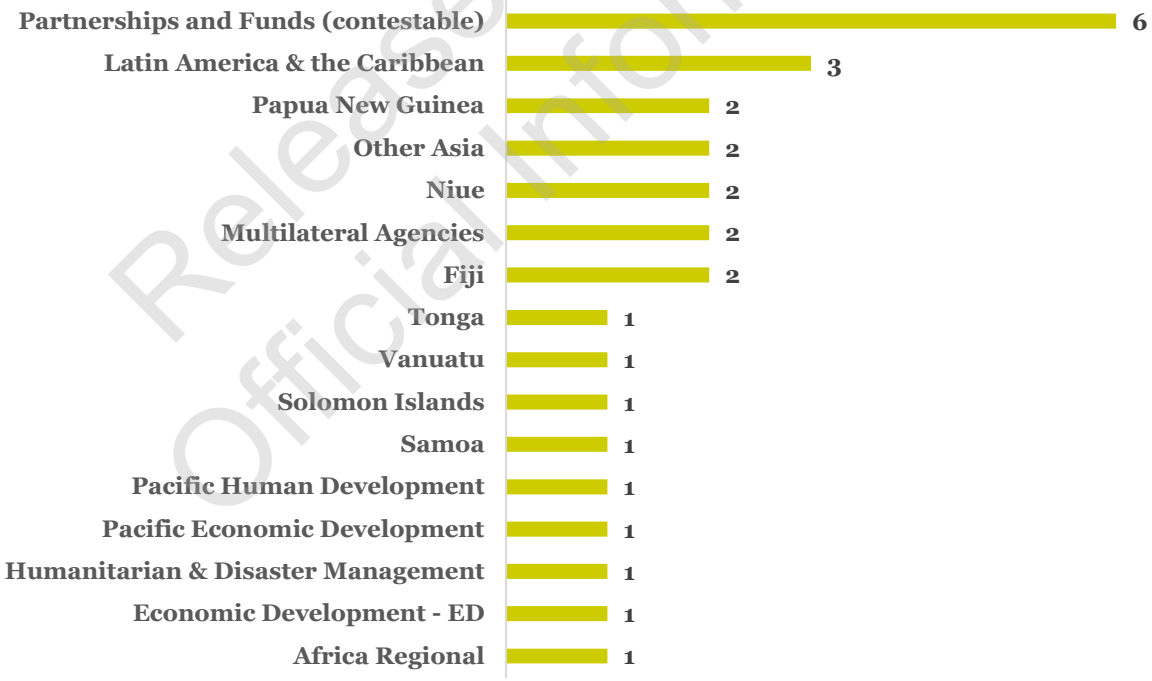
Number of ACAs by Sector



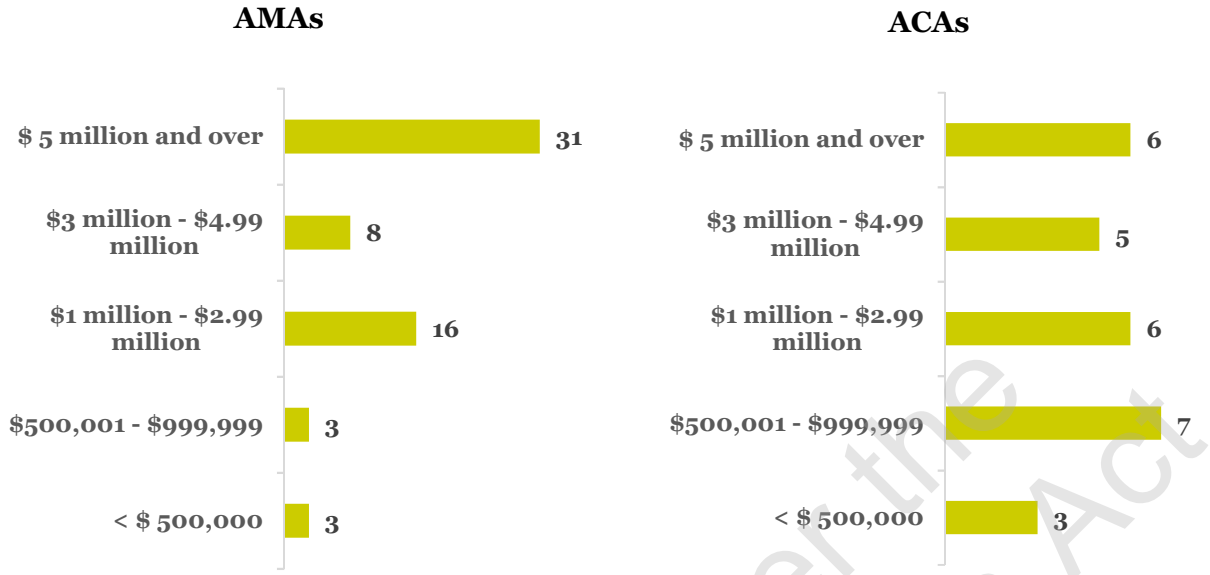
Number of AMAs by Programme



Number of ACAs by Programme



*Number of AMAs and ACAs according to Whole-of-Life Budget Programme Approval**



*The Whole-of-Life Budget Programme Approvals for three AMAs and one ACA were not available

Appendix 4: Influence of scholarship activities on AMA robustness

Robustness of effectiveness ratings in AMAs: Comparison of three AARs with scholarship activities in- and excluded

Robustness of Effectiveness Ratings (AMAs only)					
			Outputs	*Short-term outcomes*	*Medium term outcomes*
% robust post interview	**Jul 2014 – Jul 2015**	Scholarships included	73%	72%	74%
		Scholarships excluded	*84%*	*60%*	*70%*
	Jul 2016 – Jul 2017	Scholarships included	67%	62%	77%
		Scholarships excluded	*75%*	*69%*	*73%*
	Jul 2017- Jul 2018	Scholarships included	49%	56%	67%
		Scholarships excluded	*51%*	*58%*	*64%*

*Note: the difference is marginal in July 2017 – July 2018 as there are only four scholarships included in the original sample for this AAR.

Appendix 5: Assessment Templates for AMAs and ACAs

ANNUAL ASSESSMENT OF RESULTS: ASSESSMENT OF ACTIVITY MONITORING ASSESSMENTS (AMAs)

ACTIVITY INFORMATION												DETAILED REVIEW																					Overall Assessment		
												Effectiveness																					Quality of activity's Results Framework / logic and reporting and how this affects ability to track and substantiate progress and performance			Cross-Cutting Priorities			Actions to address issues (relevance, effectiveness, efficiency and sustainability)			Can Programme Managers and Deputy Directors have confidence that this AMA provides an accurate and coherent overview of the activity's progress and does it propose meaningful actions to improve the Activity?			Interview required?	
Stage of Review: 1: AAR 2: Interviews	AMS No.	Unique Identifier	Activity Name	Program	Partner	Sector / Investment Priority	Modality	Activity's whole-of-life	Old or new template	Update old or new	Technical Support	AMA Rating: Outputs	Assessment	Robust/High/Low	Notes	Code	AMA Rating: Short-Term	Assessment	Robust/High/Low	Notes	Code	AMA Rating: Medium-	Assessment	Robust/High/Low	Notes	Code	Quality of Appraisal	Notes	Code	Quality of Appraisal	Notes	Code	Quality of Appraisal	Notes	Code	Report quality	Notes	Code	Yes/No	Justification

ANNUAL ASSESSMENT OF RESULTS: ASSESSMENT OF ACTIVITY COMPLETION ASSESSMENTS (ACAs)

ACTIVITY INFORMATION											DETAILED REVIEW																																							Overall Assessment				
											Effectiveness										Relevance					Efficiency					Impact					Sustainability					Quality of activity's Results Framework / logic and reporting and how this affects ability to track and substantiate progress and performance			Cross-Cutting Priorities			Lessons Learned			Can Programme Managers and Deputy Directors have confidence that this ACA provides an accurate and reliable overview of the activity's performance and can lessons be used with confidence to inform future planning?			Interview required?	
Stage of Review: 1: AAR	AMS No.	Activity Name	Program	Partner	Sector	Modality	Activity's whole-of-life cost	Old or new template used?	Updated old new template	Technical Support Received?	ACA Rating: Outputs	Assessment	Robust/High/Low	Notes	Code	ACA Rating: Short- and Medium	Assessment	Robust/High/Low	Notes	Code	ACA Rating	Assessment	Robust/High/Low	Notes	Code	ACA Rating	Assessment	Robust/High/Low	Notes	Code	ACA rating: Impact	Assessment	Robust/High/Low	Notes	Code	ACA rating	Assessment	Robust/High/Low	Notes	Code	Quality of Appraisal	Notes	Code	Quality of Analysis	Notes	Code	Quality of Lessons	Notes	Code	Report quality	Notes	Code	Yes/No	Justification

Appendix 6: Interviewing script

INTRODUCTORY SCRIPT - MFAT Annual Assessment of Results (AAR) 2018

Thank you for agreeing to speak with me today.

As you will have been advised by the Development Strategy and Effectiveness team, I have been involved in a review of a sample of AMAs and ACAs to determine the robustness of self-assessed performance ratings in these reports. This year my colleague and I are reviewing 66 AMAs and 36 ACAs (102 in total) that were randomly selected from a total of 196 AMAs and ACAs submitted between July 2017 and July 2018.

We would specifically like to discuss the ratings given against the effectiveness criteria – these are the ratings provided for achievement of outputs, short-term outcomes and medium-term outcomes. We will aim to keep the interview to 30 minutes (longer if more than one AMA or ACA will be discussed). Time allowing, we may also briefly discuss assessments and ratings provided for other performance criteria, as well as your experience of the tools, guidance and support available to complete AMAs and ACAs.

The primary documents reviewed for our analysis are the AMAs/ACAs themselves, as well as the corresponding partner report. We assessed whether or not the effectiveness ratings given appear 'robust' on the basis of the evidence and analysis presented in the AMA/ACA and the partner report. Where further information and clarification are required, we have the option to speak with Report Authors (Activity Managers) and/or Deputy Directors in order to gain further insight into the evidence and considerations underlying the given ratings.

I asked to speak with you today because I need a deeper understanding of how you arrived at some of the ratings in the following report/s (list here)

Before we begin the discussion, I would like to assure you that the discussion will be

- Strictly confidential. Individual AMA/ACA reports will not be identifiable in the Annual Assessments of Results (AAR) report. All findings will be presented as percentages of the sample, and aggregated for smaller programs. Any comments from interviewees that are included in the report will be anonymous. Assessment and interview records will be stored securely by the Development Strategy and Effectiveness team and it will not be used for any purpose other than do inform the AAR.
- The AAR process will not result in a change to the original ratings in the MFAT system. The assessors are not tasked to recommend any changes to the ratings in individual AMAs/ACAs. Our main aim is to understand the process and quality of evidence that informed the ratings.

At the end of the interview, feedback on the independent assessment can be provided to Program Managers/Activity Managers by the interviewers, if required/requested. This feedback will be provided in general terms, that is, indicating the strengths/weaknesses of the AMA/ACA against a suite of agreed criteria against which it was assessed.

Now, to confirm, we are discussing (Activity number and name – if more than one, take them one at a time)

Questions on criteria:

- To be inserted by interviewer based on desk assessment.
- Focus on effectiveness ratings that were assessed as non-robust. Time allowing, then discuss ratings for other criteria that were also assessed as non-robust.

- Clarify evidence and interpretation of evidence to understand whether original rating is robust or not.
- In a separate row directly below the original assessment in the assessment template, indicate what the assessed rating is after interviewing and motivate why, if it has changed from the rating given at the time of the desk assessment (e.g. "The Activity Manager conducted a site visit and provided information from this visit to justify the original rating. This information was not documented in the AMA/ACA. Based on this information, the reviewer agrees with the Activity Manager's original rating – it is robust").
- If the Activity Manager cannot justify his/her original rating based on all available information, then you may want to ask about his/her interpretation of the evidence which led you to assess the original rating as non-robust. This is not about defending your assessment. It is about allowing the Activity Manager to contextualise and explain that information, which may lead you to reconsider your assessment of the original rating, or not. This is not to be debated with the Activity Manager. Simply record your final rating and the justification for it, and move on to the next item.

Questions on Process for completing AMAs/ACAs:

- Can you describe the process you went through to develop the AMA/ACA?
- Do you find the guidance and templates helpful? Any suggestions to improve them?
- Did you receive any training or support from the Insights, Monitoring & Evaluation team on how to complete AMAs/ACA? Did you find it helpful? Any suggestions to improve it?
- In your view, do you have sufficient robust evidence/data to inform the AMA/ACA process? (You could prompt about the timeliness and quality of partner reports; opportunities for site visits, etc.)
- Are there any particular aspects of AMAs and ACAs that you find especially challenging? What are they? Why do you find them challenging? (Prompt for specific aspects, e.g. mainstreaming of cross-cutting issues, VFM)