School of Global Studies, Social Science and Planning

Meta-Evaluationof AusAID's Technical Review Process

12 April 2011

Professor Patricia Rogers

Professor of Public Sector Evaluation
CIRCLE (Collaboration for Interdisciplinary Research, Consulting and Learning in Evaluation)
Centre for Applied Social Research
Royal Melbourne Institute of Technology, Australia

Summary

This report presents a meta-evaluation of AusAID's process of external technical reviews of the evaluations undertaken by external evaluators. It is based on a review of all technical reviews conducted since Jun 2008, and a review of a sample of evaluation reports, chosen to investigate the consistency of ratings, and the extent of revisions undertaken.

Technical reviews have been followed by limited improvements to draft reports

Some of the issues raised in technical reviews, such as the report structure, or lack of detail about the evidence used, have been able to be addressed in revisions to draft reports. However, based on checking changes made to low-rated reports, few revisions have been made. .

Some of the other issues raised in technical reviews have been addressed in guidance produced for evaluators and evaluation managers

The development of consistent templates and guidance has supported the improvement of the quality of evaluation reports in terms of more consistent use of key terms such as 'relevance' and 'effectiveness'. Later evaluation reports have been more likely to provide details of data sources, including lists of people interviewed. However there is still a frequent mismatch between the scope of the evaluation and the evaluator time budgeted to undertake it

Many evaluations have been rated as having insufficient data, and not reporting or using evidence well

Many evaluations have lacked suitable baseline data, quantitative data on costs, and data on intermediate outcomes and final impacts. In addition, many evaluation reports have not provided details of the methodology, including the sources of previous evaluations that have been used. Evaluations have not analysed different elements of the data to produce a stronger report, such as identifying heterogeneity of outcomes, and investigating causal attribution and contribution.

The technical review process has not reduced the frequency of identified deficiencies in terms of the evidence base for evaluations

The technical review process begins at the point where a draft evaluation report has been completed. This limits its usefulness in terms of improving individual evaluations as there is little opportunity to address gaps in the evidence underpinning the evaluation. There have been persistent problems identified in reviews in terms of the adequacy of the evidence in evaluations. In part this is caused by frequent problems with the M & E systems of the programs and projects that are being evaluated, especially in terms of developing suitable baseline data. Projects and programs need to identify from the beginning the data that should be collected to understand both what has happened after a project or outcome and also what has been the contribution of the intervention to these results.

The review process provides a consistent format but evaluations have not been rated consistently across reviewers

Overall ratings of evaluation reports have varied considerably across different reviewers. In some cases the rating seems to not be appropriate given extensive criticisms that have been made about the evaluation report.

Some form of earlier technical review is needed to ensure appropriate planning for the evaluation has been undertaken

A review of the evaluation plan (outlining the proposed methodology in detail) would increase the potential for the reviews to lead to improvements in the quality of specific evaluations.

Revised guidance on laying the foundation for a credible evaluation is needed

Most evaluations have been conducted during a short timeframe, drawing on a review of existing documents and data, key informant interviews and some field work. This makes it highly dependent on the quality of existing M & E systems and in many cases these were found to be

Background

This report presents a meta-evaluation of AusAID's process of external technical reviews of the evaluations undertaken by external evaluators. This process is complemented by an internal review process, and feedback from evaluators, which is not included in this meta-evaluation.

The procedures for the technical review process, and how these have changed over time, are set out in previous reports).

For the purposes of this report, the analysis has been undertaken in terms of three phases:

(i) June 2008 to Nov 2008 – when additional guidance documents were produced for evaluators and managers of evaluations

(ii) Dec 2008 – July 2009 – when revisions to the review process were proposed in an ODE review of the evaluation review panel

(iii) Aug 2009 – Jun 2010 – when the process was paused for a review.

Methodology

The review has been informed by discussions with AusAID staff and background reading: (i) Review of AusAID's approach to evaluation (ODE, 2007); (ii) Quality assessment and analysis of independent completion reports prepared in 2007-2008 (Kari Sann, 2008); (iii) Review of the Evaluation Review Panel (ODE, 2009); (iv) AusAID 'Rules & Tools' guidance for evaluations; and (v) a draft report prepared by Operations Policy and Support (OPS) branch (now Quality and Performance Systems (QPS) branch) in 2010 summarizing the background to the technical review process.

QPS compiled a list of 76 technical reviews and prepared a spreadsheet summarizing the quantitative data for each review. They provided copies of the technical reviews. On the basis of this data, evaluations that had been rated low in the technical review were identified, and, where possible, copies of the original report and the revised report were provided for review. The intention was not to conduct a further technical review of these, but to identify whether significant changes had been made on the basis of the technical review.

The original methodology called for reading a sample of the technical reviews. Since one of the questions related to the consistency of these reviews, and considerable variation became evident, all reviews were read.

1. What has been the value of the technical review process?

Individual evaluation reports have sometimes been revised for improved clarity

The technical review process has identified gaps and areas requiring improvement in individual reports, some of which were then addressed in revising the report. These have particularly related to improving the structuring of the report, providing a better executive summary, and providing more details about the methodology.

However there does not appear to be a systematic process for documenting which issues raised in the technical review need to be addressed in the revised report, or for archiving copies of the reports.

However, few substantial revisions have been made to evaluation reports even when they have been rated as very poor

Review of low-rated evaluation reports has shown very few revisions have been made to the draft report before it has been produced as a final report.

Serious deficiencies in the evidence base cannot be addressed by the time a draft report is produced.

The timing of the technical review process has limited its value in terms of improving individual evaluation reports. Because it has been undertaken after a draft evaluation report has been produced, there has been little scope to respond to any gaps or problems in terms of terms of reference, evaluation design (methodology), data collection or analysis. Revisions to specific evaluation reports can realistically only attend to improving the quality of reporting, such as revising the structure of the report, fine-tuning recommendations, editing the executive summary, and making more explicit the links between evidence and recommendations.

While these can improve the final report, it is too late to address gaps in data collection and retrieval that have often been identified in technical reviews.

Table 1 Consistent concerns about the adequacy of evidence presented

	Q7 Does the report present sufficient quantitative and qualitative information on costs, benefits, and performance, including baselines)?			
Date of draft evaluation report	**Agree**	**Not sure (or combination)**	**Disagree**	**% rated as satisfactory**
Jun 08- Nov 08 (Additional guidance produced)	5	2	5	42%
Dec 08 – Jul 09 ODE review of evaluation review panel)	8	4	6	45%
Aug 09 – Jun 10	21	10	14	47%

Some identified issues have been addressed in future evaluations through improved guidance and templates

The technical review process has identified some common problems that could be addressed in future evaluations by improving the guidance provided to evaluation managers and evaluators. This process has improved the quality of evaluations in relation to specific issues - in particular, using a consistent format for reports, including details of what is expected in an executive summary and the methodology section, and using a more consistent definition of relevance.

Some of the technical reviews were conducted well after an evaluation had been conducted, presumably to inform these changes, rather than to improve individual evaluations.

The process has the potential to provide clear messages to evaluators and evaluation managers about the criteria for evaluating evaluation reports

The technical review process has the potential to improve the clarity and consistency of messages to evaluators and evaluation managers about the issues that need to be addressed in evaluations, and the incentives to address these adequately. In addition to the evaluators undertaking specific evaluations, the process has provided potential evaluators who are undertaking reviews with an opportunity to become familiar with the style and focus of AusAID

evaluations. This meta-evaluation was not able to assess whether the process is being of value in this way, as it did not include interviews with evaluators or managers of evaluations.

2. What is the quality and robustness of the completed technical reviews?

Ratings of evaluation reports have been inconsistent in how they have reflected identified problems

Reviewers are asked to rate draft evaluation reports out of 5. Apart from advising that only reports scoring 3 or more will be published, no guidance is provided on the use of this scale.

There have been some inconsistencies in the ratings used, even when similar comments have been made, inconsistencies in the types of issues that have been addressed, and inconsistent understandings of key terms. This may have undermined the perceived utility of the technical review process for evaluation managers and evaluators.

In particular, it is surprising that so many evaluations were rated as having not fully addressed the terms of reference, not presenting sufficient information on costs, benefits and performance, not being sufficiently convincing in the use of evidence, not rating the intervention reasonably, and not producing a report likely to enhance AusAID's reputation for quality review and evaluation – and yet were deemed publishable reports.

Table 2 Anomalous reviews where evaluations rated publishable despite gaps in terms of key criteria

Review # and date of Technical Review	Fully addresses Terms of Reference?	Sufficient information on costs, benefits and performance	Convincing use of evidence	Reasonable ratings	Likely to enhance AusAID reputation for quality evaluation	Overall rating /5
20 (Aug 2009)	Disagree	Disagree	Disagree	Disagree	Disagree	3/5
22 (Aug 2009)		Disagree	Disagree	Disagree		3/5
30 (Oct 2009)	Not sure	Disagree	Not sure		Not sure	3/5
32 (Nov 2009)	Not sure	Disagree	Not sure	Disagree	Disagree	3/5
33 (Nov 2009)		Disagree	Not sure	Disagree	Disagree	3/5
34 (Dec 2009)	Disagree	Not sure	Not sure		Not sure	3/5
35 (Dec 2009)		Disagree	Disagree	Not sure	Not sure	3/5
37 (Dec 2009)	Not sure	Disagree	Disagree	Disagree		3/5
48 (Mar 2010)	Disagree	Disagree	Disagree		Not sure	3/5
54 (Mar 2010)	Not sure	Disagree	Not sure	Not sure	Not sure	3/5
55 (Mar 2010)	Not sure	Disagree	Disagree	Disagree	Not sure	3/5
70 (May 2010)		Disagree	Disagree	Disagree	Not sure	3/5

It is not clear why these reports were rated as acceptable. It might be that some reviewers were supportive of having evaluation reports published, and therefore used a constrained scale of 3-5 in order to advocate for publication. It might be that some reviewers rated the evaluation report in terms of its potential once identified issues were dealt with.

Several other evaluation reports have been rated highly even though there has been uncertainty about quality in terms of key criteria, as shown in the following table.

Table 3 Anomalous reviews where evaluations rated highly despite uncertainties about key criteria

Review # and date of Technical Review	Fully addresses Terms of Reference?	Sufficient information on costs, benefits and performance	Convincing use of evidence	Reasonable ratings	Likely to enhance AusAID reputation for quality evaluation	Overall rating /5
5 (Dec2008)		Not Sure			Not Sure	4/5
8 (Jan 2009)		Not Sure		Not Sure		4/5
12 (Mar 2009)		Not Sure		Not Sure		5/5
31 (Oct 2009)	Not sure	Not sure		Disagree	Not sure	4/5
36 (Dec 2009)		Not sure	Not sure		Not sure	3.5/5

3. Has the technical review process improved the quality of the evaluations over time?

Reports are now more likely to be structured and documented according to the guidelines

More recent technical reviews are less likely to identify problems with the structure of the report, or differing definitions of 'relevance'.

There are fewer unpublishable reports, but also fewer very good ones

More recent technical reviews are less likely to identify problems with the structure of the report, or differing definitions of 'relevance'.

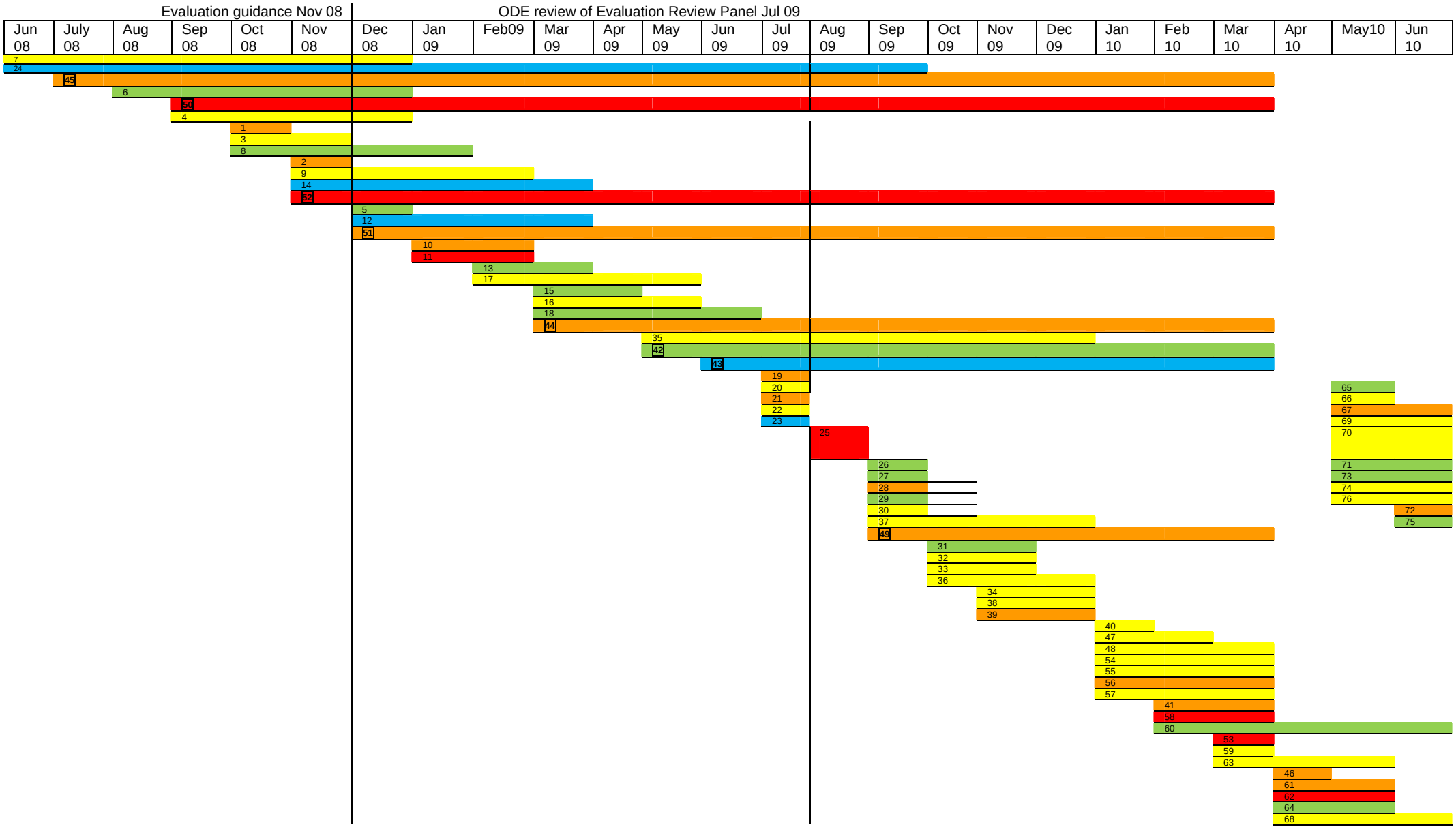
Table 4 Fewer low or high ratings of evaluation reports over time

Rating	Jun 08 – Nov 08	Dec 08 – Jul 09	Aug 09 – Jun 10
1	**2 (15%)**	**1 (5%)**	**4 (9%)**
2	**3 (23%)**	**5 (26%)**	**9 (20%)**
3	**4 (31%)**	**5 (26%)**	**21 (48%)**
4	**2 (15%)**	**5 (26%)**	**10 (23%)**
5	**2 (15%)**	**3 (16%)**	**0 (0%)**

In part this pattern is due to the increasing number of evaluation reports that are rated as a '3' (publishable), despite a list of major problems.

Table 5 Rated quality of evaluations (according to initial technical review) – rated ONE TWO THREE FOUR FIVE

Bars show the date of the draft evaluation report (beginning) and the date of the review (end) [45] Review of historic evaluation



4. Should the technical review process be continued? If yes, what should be its main purpose and how should it be improved?

The perspectives of program managers, evaluation managers and external evaluators on the value of the technical review process should also be canvassed.

The main purpose of the technical review process should be to improve the quality of evaluations undertaken.

It should do this by providing specific, relevant and timely feedback on individual evaluation plans and evaluation reports, by highlighting common issues that need to be addressed systemically, and by providing clear messages to external evaluators and evaluation managers about how evaluations will themselves be evaluated.

Based on the meta-evaluation of reviews, the technical review process should be continued and could be improved by modifying it in these ways:

Add an earlier component of the technical review, at the point where an evaluation plan has been developed.

According to the guideline “Manage the Independent Evaluation of an Aid Activity”, the evaluation plan is agreed to be the evaluation manager and the evaluation team and cleared by the evaluation delegate. The evaluation plan is not, however, appended to the evaluation reports, so it has not been possible for this review to assess these evaluation plans – in particular the extent to which they explicitly link planned data collection and analysis to evaluation questions, and whether or not gaps and deficiencies identified in the reviews would have been evident in the evaluation plan.

Some form of technical review of the evaluation plan might improve evaluations by identifying gaps and deficiencies early in the process when they can be more readily addressed, and by providing examples of evaluation plans that could inform future evaluations. There should also be provision for some flexibility in the evaluation to respond to emerging issues in field work.

Produce guidance for reviewers to improve consistency of rating.

The numerical rating of draft reports does not consistently reflect the extent or nature of concerns raised in the technical review. Evaluations where there are concerns about the adequacy of evidence, the reasonableness of how the program has been rated against evaluation criteria, and the implications for the reputation of AusAID should be clearly identified as problematic.

This guidance could include a global assessment scale for the overall rating of reports which clearly indicates how an evaluation report should be rated when it fails to adequately address major criteria, such as including adequate evidence, making defensible ratings, and being likely to add to the reputation of AusAID for quality evaluation.

More systematically and transparently identify the issues which need to be addressed in revising the draft report.

Where reviews have identified major deficiencies in evaluation reports, they should be revised unless there are compelling, and documented reasons why this is not feasible or appropriate. Identifying problems at the evaluation design stage might reduce the incidence of a dilemma where the changes needed will take significantly more time than has been budgeted for.

Feedback from management on the evaluation report is intended to be addressed by the evaluators before the report is finalized, but it is not clear if this is consistently done, nor if outstanding concerns are always documented in an appended report as outlined in the guidelines.

More regularly and systematically identify issues arising in technical reviews that cannot be addressed by revising a specific evaluation report, but which need to be addressed by revising the evaluation process.

The guidelines provide support for taking forward lessons learned in terms of programs. It would be useful to explicitly document and communicate lessons learned in terms of evaluations – both in terms of identifying problems (such as persistent gaps in baselines) and producing good evaluations (such as efficient methods for data gathering or analysis, or effective triangulation and synthesis).

1. What are the strengths and weaknesses of AusAID-funded evaluations as per the quality criteria in the technical review checklist? What are the factors affecting this?

Many evaluations have been weak in terms of evidence and conclusions and the reasonableness of ratings

The table below shows the percentage of technical reviews where it was Agreed that the evaluation had met the criterion. A high percentage of evaluation reports met criteria that related to the format of the report, the terms of reference and the issue of relevance:

- A balanced tone (Q13 – 85%)
- Clear and appropriate Terms Of Reference (Q1 – 75%)
- Being easy to read and accessible (Q13 – 75%)
- Adequately assessing relevance (Q6 – 70%)

However a low percentage of evaluation reports met the criteria (being rated either 'Disagree' or 'Not sure' or Mixed') in terms of critical criteria relating to evidence and conclusions, as well as gender analysis:

- Presenting sufficient quantitative and qualitative data (Q4 – 25%)
- Likely to maintain or enhance AusAID's reputation (Q14 – 39%)
- Gender analysis (Q5 – 42%)
- Sufficiently convincing use of evidence (Q7 – 45%)
- Reasonable evaluation criteria ratings (Q8- 45%)
- Clearly set out methodology (Q3 – 55%)
- Address Terms Of Reference (Q2- 58%)

Table 6 The percentage of evaluations rated as having met particular criteria

Q1 TOR clear	Q2 Address TOR	Q3 methodology	**Q4 Quant qual data**	**Q5 Gender analysis**	Q6 relevance	**Q7 Evidence convincing**	**Q8 ratings**	Q9 Lessons, recommendations	Q10 Exec summary	Q11 structure	Q12 Easy to read	Q13 balanced	**Q14 AusAID reputation**
75%	58%	55%	**25%**	**42%**	70%	**45%**	**45%**	68%	58%	62%	75%	85%	**39%**

Q1 Terms of Reference (TOR) for the evaluations have usually been clear but often appear unrealistic given the budget

While some earlier evaluations had unclear terms of reference, more recent evaluations have generally drawn on the templates and sample evaluation questions to produce clear and consistent terms of reference.

However the scope of the evaluations seems to be often not well matched to the evaluator time budgeted. For example, #21, the North-West Microfinance Expansion Project, had 22 evaluation questions, covering relevance, effectiveness, efficiency, impact, sustainability, gender equality, monitoring and evaluation, and lessons learned, and a time allocation of 18 days in total, including 7 days for travel and fieldwork. This would only be feasible if substantial existing data were available from previous evaluation reports and/or existing M & E systems, which does not seem to be the case for most evaluations.

This might reflect an apparent discrepancy in the expectations for evaluations. The guidelines state that an ICR (Independent Completion Report) is not the same as an Impact Evaluation, and should focus on implementation and lessons learned. However the template for an evaluation report, and the suggested evaluation questions, cover all DAC criteria, including Effectiveness and Impact.

Q2 Evaluations have often not fully addressed the TOR

Even where the technical review has identified that the TOR have not been fully addressed, this has not been changed in the final revised report. This suggests there had been an agreement to modify the TOR which had not been documented. In many cases this might have been a reasonable strategy to overcome the mismatch between the TOR and the available resources, but the changes to the TOR should have been documented and included in the report.

Q3 Evaluations have often not provided enough information about the methodology

Working within tight timelines, evaluators who have had previous experience with similar types of programs have been able to assemble and analyse considerable amounts of data quickly, and produce coherent reports, drawing on their existing knowledge. While this knowledge has been important in understanding the programs and drawing evaluative conclusions under difficult circumstances, these reports have not always been sufficiently explicit about the standards that have been used to make these conclusions, or about the evidence used.

Later evaluations have been more likely to provide detailed lists of people and groups interviewed and to provide copies of interview schedules, but few have provided details of how the data were analysed. However reports have not usually included as an appendix a matrix showing which data have been used to answer which evaluation questions, including making explicit triangulation between data.

Where evaluations have drawn on evidence from previously completed evaluations of the projects, they have not provided information about the methodology used in these.

Q5 Gender analysis has often been inadequate

In many cases this has been because data have not been collected in a disaggregated way that would allow gender analysis. There have also been reported difficulties in some cases in collecting data from women only.

Q6. Evaluation of relevance has generally been adequate

As more detailed guidance has been made available providing clear definition of what is meant by ‘relevance’, evaluations have generally addressed this criterion well.

Q7. The availability of sufficient evidence, particularly about costs and baseline data on impact variables, has been a significant weakness in many evaluations

Many reviews have expressed concern about the quality of evidence available for the evaluation - there is no baseline data available, little relevant performance information about the program’s costs, outcomes and impacts, and little opportunity for credible analysis of causal attribution and contribution.

Three major factors seem to be influencing this:

1. Many projects and programs do not have adequate monitoring and evaluation systems, and therefore important evidence is not available for the evaluation when it is conducted.

2. Most evaluations appear to be commissioned towards the end of the project or program.. The constrained time available, and the delayed start, restrict the scope for the evaluator to overcome gaps in M & E systems.

3. The methodology of most evaluations which involves a document review, briefing, field work and write up of the evaluation report in a short space of time, precludes longitudinal data collection, iterative analysis and collection, or significant triangulation and thoughtful disaggregation and analysis.

In many cases the evaluators have managed to gather and analyse an impressive amount of material in a short time frame, and may well have produced as good an evaluation report as is possible under the circumstances. It would be appropriate to consider whether it might be possible to improve the circumstances and increase their scope to produce a good report.

Q8 Ratings for the evaluation have often not been seen as reasonable

Concerns have frequently been raised about the reasonableness of the ratings. Sometimes the reviewer has advocated for a different rating on the basis of the evidence provided. Other times they have argued that while the rating might have been accurate, insufficient evidence had been presented to support it. Both cases are seriously concerning, as is the fact that there had been few revisions of the ratings in final reports.

Q9 Lessons and recommendations have generally been clear, specific, actionable and supported by the evaluation

Q10 The executive summary has usually been adequate

The executive summary is intended to be suitable for use as a stand alone document. Even after the development of a template for ICR reports, not all evaluations have met this criterion. When the technical review has rated the draft report low on this criterion it has not always been revised for the final report, even though this would be technically possible at this stage.

Q11 Most evaluations have been well structured

Where there have been concerns about the report structure, including numbering of sections, this has been addressed in revisions.

Q12 Most evaluations have been accessible and easy to read

Most reviewers have rated evaluation reports highly in terms of accessibility. One reviewer raised the issue of the level of accessibility needed – for example, being easy to read for those reading English as a second language, (including presumably international partners). Requiring all reports to be written in Plain English would increase the budget required for evaluations.

Q13 Most evaluations had a balanced tone

Fairness and balance are important criteria for an evaluation, and most reports were rated highly in this regard. However it is not clear whether reviewers were using consistent approaches to rating this. One reviewer questioned the balance of a report because all the evaluation ratings were positive, which suggests an interpretation of balance which is not about fairness but about expecting all reviews to have some low (or high) scores, regardless of performance. Some more guidance to reviewers might ensure more consistency and appropriate expectations.

Q14 Many evaluations were not seen likely to enhance or maintain AusAID's reputation for quality evaluation

The biggest concern in terms of the evaluations was how many were not seen as likely to enhance AusAID's reputation. This, together with the small scale of revisions between draft and final reports, is problematic.

Overall Rating

Meta-evaluations and subsequent revisions to reports have not been consistently conducted or archived

There appears to have been no formal tracking of evaluation report, meta-evaluations or revisions to reports. Evaluation reports have not been readily available.

2. Is evaluation quality affected by the sector, aid type, or region of the aid program being evaluated? If so, how?

There are no clear patterns in terms of quality across sector, aid type or region. Further analysis can be undertaken to see if these change when ratings are adjusted for consistency.

3. How might the quality of AusAID's evaluations be improved?

This report recommends consideration of the following revisions to the system

1. Develop reasonable expectations for evaluations
2. Improve the planning for evaluation during program planning and initial implementation
3. Extend the system to address earlier stages of an evaluation – in particular the terms of reference and the evaluation design
4. Provide additional guidance to meta-evaluators to improve the consistency of ratings
5. Develop and circulate good examples of evaluations that have succeeded in assembling and using an adequate evidence base
6. Implement clear and consistent processes for including evaluations in the system and following up the reviews in terms of implications for individual reports and for the evaluation system.

1. Develop reasonable expectations for evaluations

There needs to be clarity about what is a reasonable scope for evaluations (in terms of breadth and depth) and these need to be matched by appropriate resource allocation. At the moment evaluations have a relatively short timeline and small budget in terms of time available to collect, analyse and write up evaluation data.

The current guidelines distinguish between ICRs and Impact Evaluations; however the current Terms Of Reference usually expect ICRs to address impact.

Some options for addressing this are to:

- Clarify that ICRs should address all DAC criteria, but only to the extent possible within resource constraints
- Expect ICRs to only address impact where there is considerable existing evidence or whether a substantial evaluation budget is available

In the case of the former, it would be very useful to provide guidance and examples to show how even small scale evaluations can do more to investigate impact than report uncritically statements made by stakeholders in interviews.

2. Improve the planning for evaluation during program planning and initial implementation

The technical review process needs to be complemented by attention to developing and implementing monitoring and evaluation plans for programs and projects from commencement.

In most cases it should be possible to lay the foundation for more rigorous measurement of outcomes and impacts, including developing appropriate baselines, or capacity to reconstruct these, and for more rigorous investigation of causal attribution and contribution.

3. Extend the system to address earlier stages of an evaluation – in particular the terms of reference and the evaluation design

The technical review process could have more value in terms of improving individual evaluations if there was a process of reviewing the evaluation plan before proceeding to the evaluation – ideally at the beginning of the program to ensure appropriate evidence can be gathered.

It might be important to further develop the guidance on Terms of Reference to support different types of evaluations – in particular, it might be more useful to AusAID for some evaluations to be more in depth, and others less in depth, in order to maximise the value in terms of accountability and learning, rather than expecting all ICRs and other evaluations to have a similar scope, budget and methodological approach.

4. Provide additional guidance to technical reviewers to improve the consistency of ratings

More consistent reviews would improve their credibility and utility. Reviewers should be provided with information that describes satisfactory and unsatisfactory achievement in terms of each of the criteria, and a global assessment scale for the overall rating, which has been developed in conjunction with those who will be using the reviews.

5. Develop and circulate good examples of evaluations that have succeeded in assembling and using an adequate evidence base

A small number of diverse evaluations that have been rated highly, and been useful, should be made widely available to clearly communicate the standard required, and to demonstrate some ways of achieving this. Care should be taken to ensure that these examples cover a range of aid modalities and types of interventions wherever possible to assist evaluation managers and evaluators to learn from them for their particular evaluation.

Some of the examples should demonstrate good quantitative data analysis to address issues that cannot be addressed by simply reporting frequencies or percentages; some should demonstrate good qualitative data analysis that goes beyond providing illustrative quotes or grouping and counting comments. Some of the examples should demonstrate systematic causal analysis in terms of investigating

(i) whether the patterns of intermediate results match what would be predicted by the program theory,

(ii) whether a feasible and credible counter-factual can be developed (which might involve non-experimental methods such as process tracing or beneficiary assessment in addition to, or instead of, experimental or quasi-experimental counter-factuals such as comparison groups which may not be feasible or credible), and

(iii) whether exceptions and possible alternative explanations can be adequately ruled out.

6. Implement clear and consistent processes for including evaluations in the system and following up the reviews in terms of implications for individual reports and for the evaluation system.

The technical review process has developed over time, and is now more systematic and comprehensive. However there still does not appear to be a central information source about evaluations, their technical reviews, and the results of these technical reviews.