169 Shades of Grey: Dealing with inter-evaluator variability

Given the scope of the CEDR (more than 200 project assessments in the context of 14 country program evaluations), evidence was provided by different evaluation teams. It was important to ensure consistency in the assessment across countries and across teams, and addressing inter-evaluator variability was fundamental to the integrity of CEDR synthesis.

This article focuses on what we did to address that challenging reality, how we did it and, candidly, how we would do it differently if we had the chance again.

Penelope Jackson, African Development Bank

Introduction

Is the glass half full, or half empty?" We saw starkly the truth of that old adage recently when synthesising the results of the Comprehensive Evaluation of the Development Results (CEDR).

Addressing Inter-evaluator variability is fundamental to the integrity of CEDR synthesis and as part of this process evidence provided by different evaluation teams was brought together.

This article focuses on what we did to address that challenging reality, how we did it and, candidly, how we would do it differently if we had the chance again.

The basic building blocks of the CEDR were country evaluations, drawing on project results assessments (PRAs). Projects are the building blocks of country programs – after all the AfDB is fundamentally a project based bank. The vast majority of these PRAs were carried out within the context of the 14 country evaluations, resulting in an assessment of more than 200 projects.

The PRA approach employed goes well beyond the project completion report validation that many independent evaluation departments conduct, especially in the multilateral development bank (MDB) sphere. However, it is rarely as in depth as our full project evaluations – with few, for example, involving new beneficiary surveys. In assessing relevance, effectiveness, efficiency and sustainability, using multiple sources, it provides the right level of information to be used: both in the context of the country evaluations and for the CEDR.

Given the ultimate aim of synthesising the results of the PRAs for the CEDR, the importance of consistency in the assessment across countries and across teams was obvious from the outset. After some initial experimentation, a standard template with detailed guidance on how to assess and rate the main sub criteria, (in line with existing evaluation cooperation group (ECG) and OECD-DAC guidance) was developed for public sector operations. This was soon followed by similar guidance for private sector (or non-sovereign) operations. The latter was an interesting challenge since the difference in approach to evaluating public and private sector operations, at least amongst ECG members, is ingrained. For the purposes of the CEDR, a way to compare and collate across these two groups of operations had to be found, and one that was still aligned with existing practices and ECG guidance.

Again with consistency in mind, the PRA approach was rolled out with staff workshops to discuss each item and iron out any differences in understanding. Unfortunately, the tool was introduced too late and so, for a minority of the country evaluations, the PRAs actually had to be retrofitted with new information collected to ▸▸


PEOPLE

fill in gaps. The next batch of country evaluations were the pilots and further adjustments were made to the guidance where understanding differed across teams. The quality assurance process was piloted with 17 of the PRAs, before rolling out to the remaining 14 African countries later – including all regions, languages and development status. In sum there were more than 200 PRA available for use in the CEDR synthesis. But how far could we trust each of these in terms of rigour of the analysis, triangulation of data sources, and consistency in rating?

To be included in the synthesis database (both in terms of inclusion for the ratings and for the qualitative information) the PRAs had to go through a quality control process. In addition to ensuring the quality of the individual documents, this process was designed to ensure consistency, or at least reduce inconsistency introduced through inter evaluator variability.

At this point we expected to find little variation, perhaps a few outliers. We had, after all, followed the same detailed guidance, attended the same workshops. What we actually found was a significant degree of remaining variation between teams – both in terms of the strictness or generosity of their assessment and also the level of evidence provided to support the findings.

Why did we still see such variation? For some criteria, such as the delivery of outputs and some aspects of efficiency, the ratings are based on a numeric calculation (that is percent of planned outputs delivered with a clear rating scale showing what percent range falls within which rating). In these cases the job of the quality control is to ensure that supporting evidence can be cited, that the calculation is correct and the appropriate rating is provided as a result. However, for others there is a need for evaluator judgement, that is an assessment on a qualitative basis to inform the ratings. While this is normal in evaluation it also means that variation, particularly between evaluation teams, creeps in. Despite pages of guidance, one evaluator may see the glass as half full and another as half empty, and with a six point rating scale this can make a real difference to the aggregate results.

So how did we pick up the variation and what did we do about it? First, it was vital that we had built in a process that allowed this variation to be picked up. This also enabled us to see which criteria were subject to the highest degree of variation in the assessment – including relevance of design and matters relating to the sustainability criteria.

The process worked as follows. The country evaluation task manager submitted the PRA to the quality control process. The PRA was then reviewed by a member of the quality assurance group with no connections to the country or project in question (though where possible allocated to those with the appropriate sector expertise). During the pilot phase, two separate reviewers examined each PRA and then met to consolidate a single review. In practice, this approach was found to be too time consuming and added limited value. So, for the remainder of the PRAs, a single reviewer was allocated to each with all reviewers encouraged to exchange experiences. The dimensions of the review are included in Box 1.

The reviewer and the task manager for the PRA met to discuss the findings of the review with the former providing written comments on weaknesses and suggestions on changes, as well as a rating. Each PRA was rated in one of four categories (see Box 1). The majority of PRAs were rated in categories B or C – that is the reviewer recommended changes and in many cases stated that changes were required before the PRA could be included in the synthesis – that is to reach what the team called

Box 1: Quality assurance criteria and conclusions

The quality assurance team examined each PRA against the following dimensions:

- Evaluation design: effective use of theory of change/intervention logic
- Clarity and Rigour of Analysis: ratings and use of multiple lines of evidence
- Validity/reliability of information: data sourcing and referencing

Each PRA was then rated in one of four categories, as follows:

A. PRA meets minimum quality threshold and should be included in synthesis and no recommendations for improvement are made.

B. PRA meets minimum quality threshold and should be included in synthesis. Recommendations are made that could improve the PRA further.

C. Before including in the synthesis specific adjustments are required in order to meet the minimum quality thresholds.

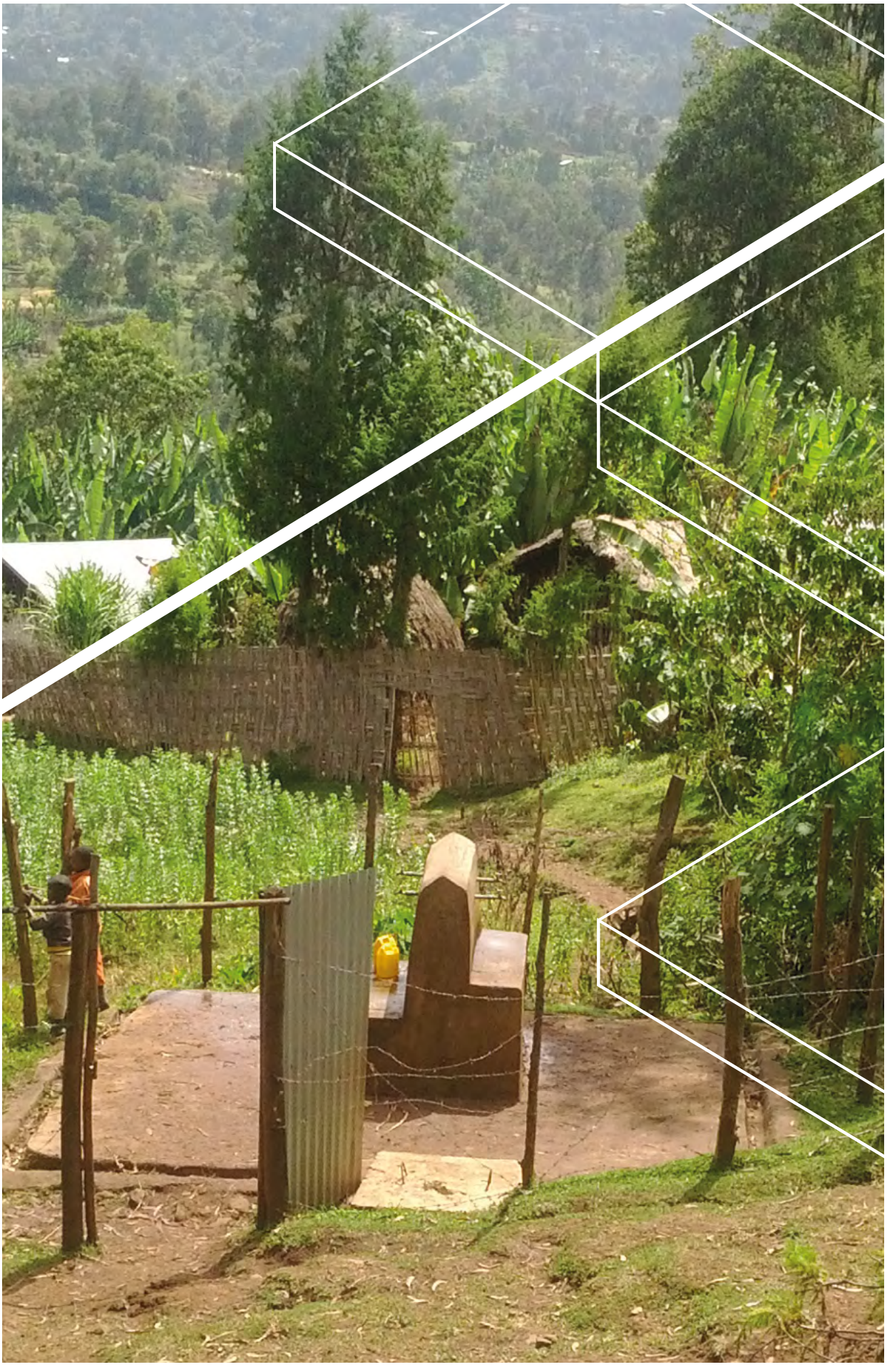
D. PRA does not meet minimum quality threshold and should be excluded from the synthesis. Significant additional work would be required to reach MQT.

minimum quality threshold (MQT). Hardly any PRAs were rated A – that is with no changes at all recommended before inclusion. However, more than 30 of the PRAs were either not provided on time or rated in category D – meaning that they should not be included in the synthesis because changes required to bring them up to standard would need to be of a fundamental nature.

Any disagreements or areas where reviewers and task managers were unclear were brought to the attention of the quality control process coordinators. The coordinators also paid close attention to PRA receiving either the top or bottom grades and the extent to which task managers effectively addressed the reviewer's comments.

At the end of the process the total number of PRA available for the final synthesis was brought down to 169. So while this reduced the statistical representativeness of the project sample available for the synthesis, this was far preferable to allowing lower quality assessments to be included.

At the end of this process, the team had a reasonable level of confidence in the comparability and quality of the PRAs included in the synthesis – and therefore in the synthesis itself. However, all involved admit the process was far from perfect and variability was never fully eliminated. In addition, the process was rushed with the team working overtime in a high stress situation. Lessons that the team drew from this experience include the following:



- **If you don't tell your teams what your minimum quality threshold is at the outset, don't be surprised if they don't all meet it.** While the PRA guidance was detailed, it would have been strengthened by clearly communicating the specific criteria against which each PRA would be quality assured, allowing teams to plan accordingly, rather than adding on later. The experience also provides lessons in what needs to be included within the PRA document, notably in relation to the explanation of the methodology used and clear referencing of sources.

> *"At the end of this process, the team had a reasonable level of confidence in the comparability and quality of the PRAs included in the synthesis – and therefore in the synthesis itself."*

- **Plan enough time not only for the reviewers to do a good job but also for the task managers to address substantive comments seriously.** The very tight time frame allowed for the quality assurance process meant that where reviewers advised teams to makes changes or find additional data, they did so on a minimal basis, that is what will allow me to meet the MQT? Time should be made to integrate significant improvements.

- **Stagger the process.** In our experience the quality assurance process started when nearly all of the PRAs had already been delivered with an extremely tight timeline for all 200 to be reviewed and revised. This had two main implications (i) it led to a large volume of work in a condensed period both for the reviewers and (ii) while some PRAs were 'hot off the press' with changes easier to incorporate, some had been completed up to three months earlier with changes harder to integrate since the results had already been taken forward in draft country strategy program evaluation (CSPE) reports.

- **Piloting – possibly the most important part of the process.** The quality assurance process was piloted with 17 of the PRAs, before rolling out to the remainder. This piloting process culminated in a full day workshop for all the reviewers to share their experiences and concerns, and for the coordinators to point to initial differences in approach. It resulted in greater progress towards consensus, revision of the quality assurance (QA) guidance and process, collective assessment of both the feasibility of the original plan (which resulted in shifting from a two reviewer to single reviewer approach) as well as their understanding of how to apply the criteria and recommendations. Following the pilot, reviewers had more confidence in their approach and work preceded more quickly.

Overall, the process was far from perfect, but it was robust enough to give IDEV staff, management and the CEDR panel of reviewers a degree of confidence in the quality and comparability of project level assessment results. The quality control effort moved us from the full spectrum of shades of grey to an acceptable range of shades within which all PRAs fell, in terms of both quality and consistency of ratings. However, we have also learned what we would do differently if we had our time again – notably building the QA criteria into the process from the beginning and allowing much more time for the process as a whole. Furthermore, it should be added that revisions to the PRAs themselves are currently being tailored for use in different types of evaluations. eVALU

Author's profile

Penelope Jackson is a Chief Evaluation Officer with Independent Development Evaluation (IDEV) of the African Development Bank. Prior to joining AfDB in February 2012, she worked for the OECD's Review and Evaluation team, where she led reviews of bilateral programmes including Japan, the EU, Portugal, South Korea, and Sweden. Her evaluation training began when she worked as the development specialist working on performance audits of the UK's Department for International Development, for the NAO. Earlier, she worked in parliamentary support in the UK and in Africa, running a range of Africa related research initiatives for parliamentary groups. Penelope has also worked in Kenya, and has led data collection missions in a many countries across the African continent and beyond. Her specialisms include robust qualitative data collection and analysis, as well as planning, and strategy development. She has a special interest in value for money and working to ensure that every development dollar is used as effectively as possible. Penelope holds an MSc (2001, Distn.) in "Violence, Conflict and Development" from the School of Oriental and African Studies (University of London).