Number 09 March 2015

Centre for Development Impact
PRACTICE PAPER
Innovation and learning in impact evaluation

Improving Quality: Current Evidence on What Affects the Quality of Commissioned Evaluations

Abstract With the increase in resources that organisations are dedicating to evaluation the issue of evaluation *quality* has risen up the agenda and a growing number of commissioners are now looking at how to ensure the studies they commission are of sufficient quality. While a plethora of evaluation quality standards exist that identify the factors that shape quality, most are experiential rather than based on research evidence. Particularly in the context of commissioning and implementing evaluation in bilateral donors, there has been limited empirical research on identifying the factors that underlie evaluation quality. Drawing on the findings of two recent studies into the quality of evaluations and other recent work in this area, this CDI Practice Paper by Rob Lloyd and Florian Schatz starts to fill this gap in evidence. The paper argues that the current debate on evaluation quality has become fixated on the issue of methodology to the neglect of other equally important issues. While methodological rigour is important, a singular focus on this issue is unwise. Considerations of quality need to permeate all stages of the evaluation process and evaluation quality needs to be recognised as a product of the capacities of the evaluation commissioner and evaluation team, the relationship between them, and the wider institutional environment in which the evaluation is being conducted.

1 Introduction

With the increase in resources that organisations are dedicating to evaluation (DFID 2014)[1] the issue of evaluation quality (see Box 1 for a definition) has risen up the agenda. A growing number of commissioners are looking at how to ensure that the studies they commission are of sufficient quality (e.g. DFAT 2014; Itad/Chr. Michelsen Institute 2014; DFID 2014; USAID 2013). For many years, the focus has been on methodology as a key factor for quality and ensuring that the methodologies used by evaluation teams are appropriate and robust. While methodology is important, there are many other factors that can shape, undermine and/or influence evaluation quality. As stated in DFID's Rapid Review of Embedding Evaluation, *'quality issues cut across most parts of the evaluation cycle, although recently the lack of attention paid to managing evaluation implementation to ensure quality has been highlighted as arguably the most critical challenge'* (DFID 2014: iii).

A plethora of evaluation quality standards recognise this and identify quality markers at all stages of the evaluation process (e.g. Yarbrough *et al.* 2011; OECD 2010; UNEG 2005; ECG 2012). However, the standards are mainly experiential rather than based on research evidence. There has been limited research on identifying the factors underlying the quality of commissioned evaluations.

Box 1 What is evaluation quality?

Quality cuts across all stages of the evaluation process. Evaluation quality includes the quality of evaluation planning and design, evaluation management, evaluation implementation, and the quality of the evaluation product itself. Existing evaluation quality standards are in line with this view and are taken as a starting point for this paper. A review of specific standards is beyond the scope of this paper.

Drawing on the findings of two recent studies of evaluation quality – for the Australian Government's Department for Foreign Affairs and Trade (DFAT) (2014) and the Norwegian Agency for Development Cooperation (Norad) (Itad/Chr. Michelsen Institute 2014) – and other recent work in this area, this CDI Practice Paper starts to fill this gap in evidence.

As space does not allow all the literature relevant to the quality of evaluations to be incorporated in this paper, it focuses on the literature from commissioning and implementation of evaluations within bilateral agencies. The rich literature generated by multilateral agencies such as the international financial institutions (IFIs) and the UN that looks at issues such as the effects of self-evaluations and decentralisation on evaluation quality is beyond the scope of this paper.

Section 2 reviews recent experience and the literature. Section 3 reveals a range of factors at all stages of the evaluation process that affect quality. It discusses the interlinkages between the identified quality factors and illustrates how the quality of the evaluation report depends on the interplay between the evaluation commissioner, evaluation team and the evaluation process managed between the two stakeholder groups. Based on the evidence presented, Section 4 provides some pointers to both evaluation commissioners and evaluators to more effectively manage evaluation quality.

2 Recent experience and literature

Within the academic literature there is limited discussion on key factors for evaluation quality, and with little primary evidence. Most recently, Cooksy and Mark (2012: 80) noted that 'achieving quality is, at least in part, an outcome of the intersection of practitioner competencies, evaluation context, and supportive resources that evaluators can access through participation in a professional community'. Other authors raise similar points, but the evidence base for these findings is largely experiential (e.g. Chelimsky 2009; Mark 2006; Wood, Apthorpe and Borton 2001).

Another source of evidence are reviews of the evaluation functions and evaluation policies of multilateral agencies (e.g. GEF 2014; UN Women 2014; WFP 2014; World Bank 2011). As mentioned above, these studies are beyond the scope of this paper.

Of the meta-evaluations that have been conducted in this area, older examples focus on assessing a sample of final evaluation reports, but with no analysis of the quality of the evaluation *process* (ALNAP 2002; Gibbons, McNally and Overman 2013; Forss *et al.* 2008). While these studies allow conclusions to be drawn about the strengths and weaknesses of the evaluation end-product, they do not help in understanding what contributed to or hindered quality. A new generation of meta-evaluations has emerged in recent years, however, that looks at the entire evaluation process and helps to advance our understanding of evaluation quality. These include studies by DFAT (2014), Norad (Itad/Chr. Michelsen Institute 2014), USAID (2013), UNDP (2013; Gariba *et al.* 2010) and ALNAP (2003, 2004). The methodology underpinning each of these reports is detailed below.

The USAID study is the widest ranging of the recent research in this area, having reviewed the quality and coverage of 340 randomly selected evaluations completed between 2009 and 2012 (USAID 2013). First, it used a quality template to review the evaluation reports and then explored the underlying factors of quality through one-on-one interviews, group discussions and a survey. Given the sample size, the study is able to correlate overall evaluation quality with different explanatory factors using chi-square and t-tests.

The study conducted for Norad was part of a larger evaluation of the Norwegian Aid Administration's approach to results measurement and evaluability. It looked at six recently completed evaluations commissioned by the Norad Evaluation Department (Itad/Chr. Michelsen Institute 2014). For each evaluation the final reports were reviewed using a standard quality assurance template. This drew from a number of existing frameworks such as the OECD/DAC quality standards for development evaluation (OECD 2011). In reconstructing the process the evaluators reviewed the original terms of reference (ToR), inception reports and draft reports, and interviewed both the evaluation managers and members of the evaluation team to gather their views of the process. By taking this approach they were able to build up the story of the evaluation.

For the DFAT study, a broadly similar approach was taken. First, the study reviewed all of the 87 operational evaluations completed in 2012 (DFAT 2014) using a quality framework based on DFAT's Aid Monitoring and Evaluation Standards. The evaluators then used statistical analysis to explore the relationship between certain explanatory factors such as sector, team composition, etc. and the quality markers. This analysis was supplemented with in-depth interviews with both evaluation managers and team members from a sample of evaluations. In selecting the sample, both high-quality and lower quality evaluation reports were selected to get a complete assessment of the enablers and inhibitors of quality.

Recent work from UNDP (2013; Gariba *et al.* 2010) and later meta-evaluations from the Active Learning Network for Accountability and Performance in Humanitarian Action (ALNAP) complement this emerging evidence base. The UNDP 2013 report draws lessons from interviews with UNDP monitoring and evaluation (M&E) advisers,

while the UNDP 2010 (Gariba *et al.* 2010) report generates findings from a quality assessment of a sample of 18 evaluations. Although undertaken some years ago, the ALNAP 2004 and 2003 meta-evaluations also moved beyond a simple analysis of evaluation report quality and included interviews with evaluation managers and evaluators as part of their methodology, allowing for some assessment of the underlying factors of evaluation quality.

3 Findings: key factors of evaluation quality

This section synthesises the findings from across the Norad, DFAT, USAID and ALNAP reports and teases out the main factors of evaluation quality. In many respects the factors discussed below are not new. Anyone involved in commissioning or undertaking evaluations will be familiar with their importance. What is interesting, however, is how frequently they are forgotten when planning and implementing evaluations – which may raise further questions about the broader system of commissioning evaluations (Winckler Andersen 2014). Part of the value of this paper is to emphasise just how important certain factors are to evaluation quality and how, in their absence, quality can be severely compromised.

Team skills The composition and skills of the team conducting the evaluation is another important factor of quality. While all of the studies point to the importance of having a generally strong team that has a balance of skills and experience between individual members, the presence of evaluation skills within the team is particularly important. The USAID (2013) study found that USAID evaluations with an evaluation specialist as part of the team were statistically of significantly higher quality (USAID 2013: 119). Similarly, interviews with DFAT evaluation managers and evaluation teams confirmed the importance of having strong *evaluation skills* within the evaluation team (DFAT 2014: 35). The Norad study highlighted similar issues (Itad/Chr. Michelsen Institute 2014: 81). Interestingly, technical expertise in a particular sector and understanding of the country or regional context were found to be of secondary importance to quality (DFAT 2014: 35). It seems that while a team requires a range of experience, there is a specific set of skills unique to evaluation that are required to deliver a quality evaluation product. The DFAT report identified technical knowledge of different evaluation methodologies; knowledge of how to lead an evaluation and the management of both international and local consultants; strong diplomatic and interpersonal skills; expertise in collecting, analysing and presenting data; and writing credible reports in a tight timescale as key evaluation skills. This is supported by other research in this area (Schwandt 2008; Stevahn *et al.* 2005; Scriven 1996).

Resourcing The meta-evaluations confirm that the level of resources for an evaluation is another key factor of quality. While none of the studies had access to comprehensive data on evaluation budgets, in the case of the DFAT study a number of proxies were used. The first proxy was 'initiative value'. Using the overall budget of the initiative that was being evaluated as a proxy for the likely evaluation budget, the evaluators found that the estimated higher evaluation budget was associated with higher evaluation quality (DFAT 2014: 30). Interestingly, this relationship only held for initiatives up to a certain value. One explanation for this could be that larger initiatives are most complex to evaluate. The second proxy used in the DFAT study was the number of evaluation days and, in particular, fieldwork days. The evaluators found a clear correlation between this and evaluation quality (*op. cit.*: 33). While the USAID meta-evaluation did not have access to reliable data on evaluation budgets or duration and therefore could not test the association between these factors and evaluation quality, interviews with USAID evaluators identified time as a key quality factor (USAID 2013: 10). A similar finding emerged from the interviews conducted for the ALNAP study (ALNAP 2004: 144), interviews with UNDP M&E advisers (2013: 40) and a quality assessment of UNDP evaluations (Gariba *et al.* 2010: 32). A clear implication of this finding is the importance of commissioners costing evaluations appropriately.

Purpose The extent to which an evaluation has a clear purpose is strongly correlated with evaluation quality. This was a common finding across the DFAT, USAID and ALNAP studies. When there is clarity around why an evaluation has been commissioned and how it is going to be used, quality seems to be higher. Evaluations that have a *clear purpose to inform management decisions*, in particular on future programming, are correlated with higher quality (DFAT 2014: 31; USAID 2013: 119; ALNAP 2004: 133). While none of the studies investigate the underlying reasons for why purpose is so important, we would argue that having clarity around how an evaluation is going to be used means people are more invested in the process, are more likely to monitor and help shape quality and are therefore ultimately more likely to use the findings. On the other hand, the movement towards independent evaluation raises the issue of interference when evaluations are too close to management.

Planning Interviews with evaluation managers and evaluation teams at DFAT suggested that *allowing sufficient time for planning* was central to evaluation quality. It was mentioned that evaluations need to be planned at least six months in advance so that the best evaluators are contracted, country visits are well prepared and meetings with the right stakeholders set up, and adequate time is available for preparation and report writing (DFAT 2014: 34). It was felt that with a shorter planning period the time frames became too compressed, too many compromises were made and quality slipped as a consequence. In the case of the ALNAP study, evaluation managers indicated that at least three months of preparation are needed to

identify suitable evaluators and agree the terms of reference (ToR). It was noted that *'more time and effort before an evaluation pays back ten times'* (ALNAP 2004: 144).

Number of evaluation questions There is some evidence that the number of evaluation questions is related to its quality. The Norad study found that evaluation ToRs asked 22–29 evaluation questions, which spread resources too thinly and allowed evaluators to focus on the questions that are easier to answer.[2] Evaluations with too many questions are unlikely to generate in-depth analysis and to document impact – which is often the more challenging part of an evaluation (Itad/Chr. Michelsen Institute 2014: 72). About half of the evaluation managers interviewed for the ALNAP study felt that the ToRs were overloaded with questions (ALNAP 2004: 136). USAID evaluation managers and evaluators also felt that a large number of evaluation questions could impede evaluation quality, yet no statistically significant correlation between the number of evaluation questions and evaluation quality could be found (USAID 2013: 27, 123). DFAT evaluators noted cases where the overall scope of the evaluation, including the number of evaluation questions, was too ambitious in relation to the budget and as a result quality suffered (DFAT 2014: 37). Interestingly, the number of evaluation questions relates strongly to the capacity of the commissioner (see below); however, this link was not made explicit in any of the studies.

Commissioner capacity There is evidence that capacity gaps among evaluation commissioners can negatively affect evaluation quality. Interviews with evaluation managers and evaluation teams highlighted the effect that DFAT staff not having the time or the skills to manage evaluations can have on evaluation quality. A common issue that emerged was that evaluation skills at DFAT are stretched and sometimes less experienced staff are tasked to manage evaluations (DFAT 2014: 38). One of the contributing factors to this was DFAT's evaluation policy at the time that made programme evaluations mandatory, which meant a large number of staff had to commission and manage evaluations as part of their programme management duties.[3] In the case of USAID, evaluation providers commented that a lack of evaluation skills among commissioners manifested in poor ToR which they found difficult to respond to (USAID 2013: 111). Interviews with Norad staff suggested that the wide range of evaluations commissioned, both methodologically and thematically, was possibly too demanding for commissioner staff and has an impact on quality. Staff noted that quality was particularly difficult to manage when evaluations used methodologies of which they had no experience (Itad/Chr. Michelsen Institute 2014: 72).

Institutional factors There is evidence that institutional pressures and policies affect evaluation quality. These factors play a role early on but continue to influence quality throughout the evaluation process. For instance, interviews with both evaluation managers and evaluators at DFAT suggested that the evaluation policy had perverse effects on quality. They argued that the *drive for increasing the number of evaluations* risked promoting a compliance-driven approach in evaluations (DFAT 2014: 38–9). If evaluation managers are pressured to commission more evaluations than they have time to manage, evaluations are likely to become a 'tick-box' exercise. In this type of environment quality inevitably dips. Conversely, in the case of USAID, there was a clear correlation between evaluation quality and whether an evaluation was commissioned before or after the introduction of new *evaluation policies and systems*. The changes included the strengthening of M&E as one of the topline performance indicators of reform efforts, the roll-out of evaluation training courses to 1,200 USAID staff members and other stakeholders, and the introduction of a target for high-quality evaluations through the Forward Initiative (USAID 2013: 2). Interviews with evaluation managers and evaluators confirmed the influence of the new evaluation policies and systems on quality (USAID 2013: 120). Evaluation managers interviewed for an ALNAP meta-evaluation stressed the importance of internal *buy-in and ownership* as another key institutional factor influencing evaluation quality (ALNAP 2004: 133, 170). Interviews with UNDP M&E advisers confirmed that the increasing demand for evaluative evidence by senior management has been a critical factor in improving the quality of evaluations (UNDP 2013: 40).

While the evidence points clearly to the importance of the wider enabling environment for understanding evaluation quality, it also suggests that the relationship between the two variables is complex. In the case of USAID, the introduction of the new evaluation policy had a direct effect on driving up evaluation quality; in the case of DFAT the evaluation policy seemed to reduce the quality of evaluations because staff lacked the skills to cope with the additional demands of commissioning and managing more evaluations.

M&E system There is some evidence that poor quality monitoring data affects evaluation quality. While an absence of data meant that the DFAT study was not able to establish the precise nature and extent of the relationship between the quality of the monitoring system and evaluation quality, there was sufficient evidence both from the statistical analysis and interviews with evaluation managers and evaluation teams to suggest a relationship (DFAT 2014: 31). When the performance management system is of poor quality, either in terms of what data are being collected or how they are being collected, the evaluation also tends to be of low quality. A similar finding emerged from the Norad study. The absence of initiative-level monitoring data is one of the main reasons for poor evaluation quality and why it proved difficult to demonstrate the difference that Norwegian aid makes

(Itad/Chr. Michelsen Institute 2014: xvii, 88ff). An assessment of UNDP evaluations confirmed the role of reporting and monitoring in influencing the quality of evaluations (Gariba *et al.* 2010: 32).

Communication Interviews with DFAT evaluation managers and evaluators emphasised good communication between the evaluation manager and the evaluation team as a key quality factor. *Mutual respect, learning and transparency* all appear to strengthen evaluation quality. For instance, there were positive examples where evaluators were able to fine-tune ToRs in consultation with evaluation managers, allowing them to have a better understanding of the key objectives of an evaluation from the beginning. Similarly, evaluators felt that it helped them to improve the quality of future evaluations when they were invited to do follow-up work and see how their evaluation reports were utilised. Also, negative examples were noted where evaluation managers were not transparent about the purpose of an evaluation (DFAT 2014: 35). In the case of Norad, evaluation managers indicated that there was limited communication between them and the evaluation teams. It was found that this *hands-off approach* can affect evaluation quality because evaluation managers get sight of a report only after data have already been collected and analysed, too late in the process to enact any significant changes (Itad/Chr. Michelsen Institute 2014: 73).

4 Conclusions

While the factors of quality of evaluation discussed above are not new – they will be familiar to any evaluation commissioner or evaluator – it is surprising how often they are forgotten (or are undervalued) in the planning and implementation of an evaluation. Part of the problem is that the debate around quality of evaluation gravitates towards methodology and the end-product, i.e. the final evaluation report. The evidence from recent studies of quality in the evaluation process indicates that a rethink is required. While issues such as methodological rigour are important, a singular focus on this (or any other single quality factor) is unwise. Considerations of quality need to permeate all stages of the evaluation process. Evaluation quality is a product of the capacities of the evaluation commissioner and evaluation team, the relationship between them, and the wider institutional environment in which the evaluation is being conducted (Winckler Andersen 2014). With the increase in resources being put into evaluation, broadening the approach to quality of the evaluation process is essential. If we fail to do so, the recent gains that have been made in promoting evaluation use, and evidence more broadly, in policymaking may be undone. In the face of low-quality evaluation reports decision-makers will question the utility of investing in the generation of evaluations and the current enthusiasm for them, and it may be that the associated resourcing of them, will reduce.

This paper is only the start of an exploration of the factors that influence the quality of evaluations. More research is needed to fully understand how these factors interact with each other and through which channels they affect overall evaluation quality. The relative importance of quality factors also remains unknown. Existing literature from multilateral reviews of evaluation functions and evaluation policies can provide more insights and need to be linked to the emerging body of evidence from meta-evaluations. We would suggest that developing a deeper understanding in this area is important to evaluation commissioners and evaluators alike.

Notes

1 While this paper is looking at all types of evaluation, irrespective of whether they are impact evaluation or not, the findings are just as much applicable to impact evaluations as others.

2 See also Winckler Andersen (2014) and Barnett and Bennett (2014).

3 In early 2012, the department's evaluation policy was revised, reducing the number of mandatory operational evaluations by approximately half to only one during the lifetime of each aid initiative. Further changes planned for mid-2014 will raise the financial threshold for aid initiatives requiring mandatory evaluation and will reduce evaluation numbers by a further 42 per cent.

References

ALNAP (2004) *Review of Humanitarian Action in 2004*, Active Learning Network for Accountability and Performance in Humanitarian Action, London: Overseas Development Institute

ALNAP (2003) *Review of Humanitarian Action in 2003*, Active Learning Network for Accountability and Performance in Humanitarian Action, London: Overseas Development Institute

ALNAP (2002) *Review of Humanitarian Action in 2002*, Active Learning Network for Accountability and Performance in Humanitarian Action, London: Overseas Development Institute

Barnett, C. and Bennett, J. (2014) 'Critical Reflections on the South Sudan Evaluation of Conflict and Peacebuilding Activities', in O. Winckler Andersen, B. Bull and M. Kennedy-Chouane (eds), *Evaluation Methodologies for Aid in Conflict*, London: Taylor & Francis

Chelimsky, E. (2009) 'Integrating Evaluation Units into the Political Environment of Government: The Role of Evaluation Policy', in William M.K. Trochim, Melvin M. Mark and Leslie J. Cooksy, *Evaluation Policy and Evaluation Practice. New Directions for Evaluation* 123: 51–66

Cooksy, L.J. and Mark, M.M. (2012) 'Influences on Evaluation Quality', *American Journal of Evaluation* 33.1: 79–84

DFAT (2014) *Quality of Australian Aid Operational Evaluations*, June, Canberra: Office of Development Effectiveness, Department of Foreign Affairs and Trade, Australian Government

DFID (2014) 'Rapid Review of Embedding Evaluation in UK Department for International Development', February

ECG (2012) 'Big Book on Evaluation Good Practice Standards', Evaluation Cooperation Group

Forss, K.; Vedung, E.; Kruse, S.E.; Mwaiselage, A. and Nilsdotter, A. (2008) *Are Sida Evaluations Good Enough? An Assessment of 34 Evaluation Reports*, Sida Studies in Evaluation 2008: 1, Stockholm: Sida

Gariba, S.; Balogun, P.; Thanh An, P.T. with Hildenwall, V. (2010) *Independent Review of the UNDP, Evaluation Policy*, New York: UNDP

GEF (2014) *Report of the Second Professional Peer Review of the GEF Evaluation Function*, Washington DC: Global Environment Facility Independent Evaluation Group

Gibbons, S.; McNally, S. and Overman, H. (2013) 'Review of Government Evaluations. A Report for the NAO', prepared by London School of Economics (LSE) for National Audit Office

Itad/Chr. Michelsen Institute (2014) *Can We Demonstrate the Difference that Norwegian Aid Makes? Evaluation of Results Measurement and How this can be Improved*, Oslo: Norad

Mark, M. (2006) 'The Consequences of Evaluation: Theory, Research, and Practice', Presidential address presented at the annual meeting of the American Evaluation Association, St Louis, Missouri

OECD (2010) *Quality Standards for Development Evaluation*, Development Assistance Committee Guidelines and Reference Series, Paris: OECD Publishing

Schwandt, T.A. (2008) 'Educating for Intelligent Belief in Evaluation', *American Journal of Evaluation* 29: 139–50

Scriven, M. (1996) 'Types of Evaluation and Types of Evaluator', *American Journal of Evaluation* 17: 151–61

Stevahn, L.; King, J.A.; Ghere, G. and Minnema, J. (2005) 'Establishing Essential Competencies for Program Evaluators', *American Journal of Evaluation* 26: 43–59

UNDP (2013) *Annual Report on Evaluation 2013*, New York: Independent Evaluation Office, UNDP

UNEG (2005) *Standards for Evaluation in the UN System*, Foundation Document, New York: United Nations Evaluations Group

UN Women (2014) 'Professional Peer Review of the Evaluation Function of UN Women', September

USAID (2013) *Meta-evaluation of Quality and Coverage of USAID Evaluations 2009–2012*, prepared by Management Systems International

WFP (2014) 'Peer Review of the Evaluation Function of the UN World Food Programme (2008–2013)', prepared by OECD Development Assistance Community – UN Evaluation Group

Winckler Andersen, O. (2014) 'Some Thoughts on Development Evaluation Processes', *IDS Bulletin* 45.6: 77–84

Wood, A.; Apthorpe, R. and Borton, J. (2001) *Evaluating International Humanitarian Action*, London: Zed Books

World Bank (2011) 'Self-Evaluation of the Independent Evaluation Group', World Bank Independent Evaluation Group

Yarbrough, D.B.; Shulha, L.M.; Hopson, R.K. and Caruthers, F.A. (2011) *The Program Evaluation Standards: A Guide for Evaluators and Evaluation Users* (3rd ed.), Thousand Oaks CA: Sage

Centre for Development Impact (CDI)

The Centre is a collaboration between IDS (www.ids.ac.uk) and ITAD (www.itad.com).

The Centre aims to contribute to innovation and excellence in the areas of impact assessment, evaluation and learning in development. The Centre's work is presently focused on:

(1) Exploring a broader range of evaluation designs and methods, and approaches to causal inference.
(2) Designing appropriate ways to assess the impact of complex interventions in challenging contexts.
(3) Better understanding the political dynamics and other factors in the evaluation process, including the use of evaluation evidence.

This CDI Practice Paper was written by **Rob Lloyd** and **Florian Schatz**.

ISSN: 2053-0536
AG Level 2 Output ID: 315

Institute of Development Studies, Brighton BN1 9RE, UK
T +44 (0) 1273 915637 **F** +44 (0) 1273 621202 **E** ids@ids.ac.uk **W** www.ids.ac.uk