

***DAC Evaluation Network: Follow-up to
the Paris Declaration on Aid
Effectiveness***

Review of Literature

To accompany the Options Paper
revised after discussion at the fourth meeting of the
DAC Evaluation Network, 30-31 March 2006

by

Alison Evans
*Institute of Development Studies,
University of Sussex*

David Booth
*Overseas Development Institute,
London*

**Revised Draft
4 May 2006**

Please send any comments to
d.booth@odi.org.uk; a.evans@ids.ac.uk;

Contents

1	Introduction	1
2	The challenge	1
2.1	Linking aid effectiveness and development effectiveness	1
2.2	Evaluation frameworks and the meaning of theory-based evaluation	3
2.3	Evaluating change processes and evaluating impacts	6
3	Design of the evaluation work	9
3.1	Basic design issues	9
3.2	Objectives and trade-offs	13
	<i>References</i>	16

1 Introduction

This draft review of literature has been prepared to support the discussion in the DAC Network on Development Evaluation on ways of following up the Paris Declaration on Aid Effectiveness. It is designed as a focused literature review to accompany the specific points in the Options Paper submitted for discussion at the March 2006 meeting of the Network and subsequently revised. It provides additional examples and illustrations from published literature and recent unpublished documents to substantiate some of the points made in the Options Paper. It is not intended as an exhaustive review of all available literature on aid effectiveness and evaluation and it will be updated after the Evaluation Network meeting in view of any additional documents provided by members or partners.

For ease of reference, the Literature Review uses the same main section headings as the Options Paper. Specific references to the Literature Review have also been included at key points in the Options Paper.

2 The challenge

2.1 Linking aid effectiveness and development effectiveness: Lessons from World Bank Annual Reviews of Development Effectiveness

One of the most challenging features of the Paris Declaration from the point of view of evaluation is the way it links aid effectiveness with development effectiveness. The Declaration asserts not only that the actions to be undertaken on Ownership, Alignment, Harmonisation, Management for Results and Mutual Accountability will result in more effective aid, but also that this will contribute to countries' development performance. The World Bank OED's *Annual Reviews of Development Effectiveness* (ARDEs) are a rich source of insight into the relationship between effective aid delivery and "project/programme outcomes". We examined the full run of ARDEs (1997-2005) for clues as to the challenges likely to be faced by evaluation work spanning aid effectiveness and development effectiveness.

2.1.1 World Bank ARDEs

Themes recurring in the ARDEs include:

- ❑ ***The central role played by institutions***, with clear evidence of a positive association (but not causation) between measures of institutional quality and project and policy performance.
- ❑ The institutional collateral damage resulting from aid channelled through isolated, uncoordinated enclave projects.

- ❑ The need for consensus-building (and the time needed for that) as well as respect for partner country ownership and leadership, and the need to recognise **possible contradictions or tensions within the aid effectiveness paradigm** – e.g. between donor demands for increased financial accountability and calls for a country-led partnership.
- ❑ The importance of adapting strategies to country context, including a blend of instruments that can be flexibly attuned to country priorities.
- ❑ The need to balance *comprehensivity* (especially in analysis and understanding) and *selectivity* in action.
- ❑ The need to integrate a learning process into programming (hence the use of adaptable lending instruments such as APLs and LILs), to take evaluable/measurable steps, and to engage in joint monitoring as far as possible.

The causal or logic chain linking effective aid delivery and successful project outcomes is implicit rather than explicit in ARDEs, although it is evident that much rests on the quality of Bank inputs and the quality of receiving institutions in partner countries.

2.1.2 Association and causality in institutional analysis

Successive ARDEs focus on understanding how the way in which the Bank does business is effective in supporting institutional change, or not. The argument is that effective institutions deliver better “project outcomes”; and presumably that means better “development outcomes”. It remains unclear, however, how much of this belief in the centrality of institutions is theory-based or based on a series of associations found in cross-country regressions. The fact that there is a positive association between measures of institutional quality and Bank project success is perhaps not surprising, but which way the causation runs remains unanswered. ***There is relatively little theory to help untangle which way the causalities run and why.***

Some other observations from the ARDEs would be important for any evaluation of the Paris Declaration:

- ❑ ***The total impact of donor efforts at country level may be more than, or less than, the sum of its parts.*** This raises the possibility that in some country contexts even a limited number of behaviour/procedural changes could make a huge difference to the way partner governments do business and the size and effectiveness of their resource envelope. However, in other country settings, even with considerable donor effort, development outcomes might not change, or might change very little, because of a range of structural/exogenous factors.

Understanding the interaction between country conditions and the prospect of converting better aid into better development outcomes could be a crucial area of added value for evaluation of the Paris Declaration.

- ❑ ***Demonstrating attribution is a difficult process for any evaluation but more so when the implementing environment is so complex.***

This suggests the value of selecting out specific evaluation fields for in-depth study, so that some at least of the complexities can be managed. For example looking at how donor behaviour (across the range of Paris Declaration commitments) is changing in a specific sector, or focusing on one particular dimension of the Paris Declaration, such as alignment or results management.

It also points to the potential value of moving towards a modified notion of attribution, what the Bank now calls “most likely association”. This allows for a sound evaluative judgement based on the best evidence available while at the same time acknowledging that conditions are far from experimental and that data and knowledge gaps are widespread.

2.2 Evaluation frameworks and the meaning of theory-based evaluation

Particularly in fields where there is little formal theory and there are substantial knowledge gaps, it is important to be clear about the purpose of constructing an evaluation framework. Characterising the approach to be adopted as “theory based” is likely to lead to misunderstanding and to unnecessary objections if the meaning of that phrase is not well explained. Weiss (1998) is a classic source on the subject. A discussion of the paper on evaluating aid harmonisation by Balogun (2005) and the draft report of the General Budget Support Evaluation (IDD and Associates, 2006) helps to illustrate some of the main points.

2.2.1 Weiss on theory-based evaluation

Writing originally in the early 1970s, Weiss observes that many evaluations have been undertaken by investigating outcomes without much attention to the paths by which they were produced. These days, however, evaluation questions typically include not just “did the programme work?” but “what made it work?” and even “how can we make it better”? In order to contribute in a serious way to such a discussion, it helps if the evaluator understands, and investigates, the programme’s explicit or implicit theory.

Theory in this context refers not to the kind of formalised body of knowledge that one encounters in natural or human sciences, but sets of stakeholder beliefs about how programme objectives are likely to be achieved. There is no implication that the “programme theory” is right or that there is a consensus about it. It does not have to be based on systematic research, although it may be at least partially so (1998: 55, 61-62).

So what is the point of attempting to lay out a programme theory as the basis of an evaluation framework? It is to allow the evaluator to organise questions and evidence-based enquiries in such a way that they trace the unfolding of the assumptions that gave rise to the programme’s design. This makes it easier to

distinguish unintended consequences and unexpected blockages. ***It makes it more likely that the findings of the evaluation will feed back into the general thinking that inspired the programme design, and thereby contribute to the generation of more effective programmes in the future.*** Meta-analysis, which draws together findings on a large numbers of programmes of the same type, may be particularly well placed to do this (1998: 60-61, 69). Needless to say, all of these points apply to policy commitments that are based on a certain set of ideas about how objectives are likely to be met, as well as to programmes as such.

2.2.2 Balogun on evaluating harmonisation

Paul Balogun's (2005) study for DFID on evaluating progress towards harmonisation discusses several of the issues confronted in our Options Paper, but treats the question of "theory" differently from the way we would recommend. Focusing on aid harmonisation, Balogun stresses the importance of defining an evaluation framework that identifies the contribution of harmonisation to enhancing the efficiency and effectiveness of the overall aid management system at country level. However he finds that "it is impossible at present to develop a usable evaluation framework to answer these questions" (Balogun, 2005).

According to Balogun, much of the difficulty stems from the fact that an evaluation framework assumes that a *deductive* evaluation approach can be used. That is:

- "that the evaluation framework, based on either theory or empirical evidence, clearly predefines the key actions necessary to efficiently and effectively deliver the desired impact; [and]
- that the key contextual factors which determine what can be done and what will be most efficient and effective, given the particular circumstances, are known, understood and reflected in the framework".

However, "At present, neither the theory nor the empirical evidence necessary to meet either of these assumptions is available" (Balogun, 2005: S16-17). Balogun therefore argues that there is a need to start with an *inductive* approach as a basis for amassing the necessary empirical evidence upon which to strengthen the evaluation framework.

Balogun does, nevertheless, consider it is possible in the short term to identify a number of reasonable hypotheses given our understanding of how organisations and institutions operate. Such hypotheses can be used to frame discussion of the possible benefits of harmonisation and to advance stakeholders' understanding of harmonisation and how it may contribute to aid effectiveness. This is arguably the basis for a formative evaluation.

Balogun is undoubtedly right about the challenges involved in evaluating harmonisation, and the challenges are multiplied if all five of the areas covered by the Paris Declaration are to be investigated simultaneously. However, the

suggestion that constructing an evaluation framework implies a “deductive approach” seems wrong. We see no difference between Weiss’s characterisation of a theory-based evaluation approach and the actual procedure Balogun proposes of framing discussion within a set of “reasonable hypotheses”.

2.2.3 *Negative findings in theory-based evaluation: the GBS experience*

For the reasons just explained, evaluation literature does not expect that “programme theories” will be robust *ex ante*. Indeed, if programme theories were known to be entirely sound, there would be little need for evaluation. ***It follows that evaluation findings that show up the weaknesses of the relevant programme theory are at least as useful as those that indicate their strengths.*** For this reason, a good evaluation framework will normally include evaluation questions signalling possible ways in which the expected causal linkages might *not* be working, including the intervention of exogenous factors (equivalent to the Risks and Assumptions column of a Logical Framework).

In practice evaluators sometimes underestimate the value of negative findings. This is illustrated by the recently completed General Budget Support evaluation.

In their draft synthesis, the GBS evaluation team report: “Some commentators have remarked that some of the potential effects depicted in the causality map – particularly those relating to changes in basic institutional relationships – are implausible in terms of the transformational and political character of the effects that are sought” (IDD and Associates, 2006: 102). This comment is entirely consistent with the initial framework for the GBS evaluation, which set out the assumptions about domestically-driven institutional changes that were necessary to make plausible the expectation of positive institutional effects from GBS inputs. Test questions were written into the framework on these points – “Are there factors in the country’s changing political climate that are raising demands for greater accountability to domestic actors”, etc. (Lawson and Booth, 2004: 47, 59). Emerging experience from the GBS evaluation implies that there are real challenges in identifying negative causality flows as distinct from positive or neutral findings about the programme theory generally.

It would be important for evaluation work on the Paris Declaration to take very seriously the equivalent set of issues. These are to a larger extent internal to the basic results chain, as this encompasses both country institutional issues and donor behaviour. But “exogenous” factors, including aspects of entrenched donor behaviour that were *not* the subject of Paris Declaration commitments would also be potentially important (Rogerson, 2005).

2.3 Evaluating change processes and evaluating impacts

Where one is interested, as we are here, in results chains that go all the way from behaviours in the aid relationship to poverty impacts, at least three large questions pose themselves:

- ❑ In view of the difficulties laid out by Balogun, is there a way simplifying the task of evaluating complex changes in the aid relationship, to make them more manageable?
- ❑ How far down the results chain starting with Inputs and Outputs should the evaluation be expected to go in investigating causes and effects, or “plausible association”?
- ❑ What is the bearing on all of this of the increasingly compelling argument that there is a major “evaluation gap” at the level of impact evaluation?

We consider the most salient literature on these points.

2.3.1 *Simplifying complex institutional change*

One way of responding to the challenges posed by the evaluation of large-scale institutional change processes is to relax somewhat the degree of specification of the “theory” being tested. In an original approach Elbadawi and Randa (n.d.) examine the development impact of the Comprehensive Development Framework by analysing the experiences of countries that have adopted development strategies that *approximate the CDF concept of development*, or what they term *CDF-like development strategies*.

To give empirical content to the concept, they develop parallel indicators of *CDF-like* principles which in turn are aggregated to form an overall *CDF-like* index. This index is subsequently combined with other global development data to analyse the development impact of *CDF-like* development experiences.

The key evaluation question is whether *CDF-like* development strategies have helped to produce better development outcomes, including better institutions, higher growth, lower poverty and enhanced human development, as well as contributing to improved aid effectiveness. This makes it quite similar to the central question examined in the Options Paper. There are two main causal hypotheses: that a *CDF-like* development strategy is a function of good policy and institutions (or vice versa) or, alternatively, that no causation exists and that both are driven by deeper country-specific characteristics such as the initial level of development, the degree of social or economic fractionalisation etc.

A key finding is that countries with higher income, functioning democracy and less fractionalised societies are more likely to adopt *CDF-like* strategies as well as to have better institutions. The logic is that *CDF-like* development strategy is a process that should influence better intermediate outcomes, such as institutions. Hence, initial conditions influence institutions *through* their influence on development processes and strategies. By using a control-group model, Elbadawi et al. also find that *CDF-like* development strategies have had a strong marginal contribution to growth, literacy and institutions. Aid, on the other hand, does not have a direct impact on growth or institutions. Instead aid is more effective where a *CDF-like* development strategy exists.

Implications for the evaluation of the Paris Declaration are several:

- *Paris-Declaration-like* behaviours could be the focus of the evaluation in the first instance.
- Initial conditions are likely to matter both in terms of those countries likely to adopt the Paris Declaration commitments and the effect that Paris Declaration commitments have on institutions and, therefore, on development outcomes.
- Specification of a control group would facilitate an examination of the marginal contribution of Paris-Declaration-like changes.
- Aid enhances growth through its influence on development processes and strategy; hence a key link between aid and development effectiveness is the support that aid gives to broad political and social processes and not just the institutions of development policy (Elbadawi and Randa, n.d.).

2.3.2 *How far down the results chain?*

The GBS evaluation is a fair guide to the maximum that may be possible in evaluating the Paris Declaration commitments, taking into account the Declaration's large scope. The GBS evaluation worked with an evaluation framework that was as comprehensive in respect of results levels as the outline framework for the Paris Declaration in the Options Paper. The evaluation team found this chain of causality to be long, both conceptually and temporally:

Following a results chain all the way from inputs to impact is known to be challenging, particularly in moving from outputs to outcomes and impact. In any circumstances, the intervals and their immediate effects and outputs, outcomes and impacts will be significant. When effects are expected to result from processes of institutional change, the plausible interval for effects to be manifested is longer still. Moreover, results may be such that they are measurable only periodically and with difficulty; this lengthens the interval, in practice, before results can be ascertained. In some cases, moreover, confidence in the reliability of a link from hypothetical causes to observed effects may require repeated observations and evidence that the effect is persistent (IDD and Associates, 2006: 9-10).

The difficulties were particularly acute for the country evaluations in Malawi and Nicaragua, where budget support had only recently come on stream. But they were serious enough even in countries like Uganda with more than five years of experience with the modality. Similar problems confronted the independent evaluation using the same basic framework in Tanzania, even though something approaching a decade of experience was examined (Lawson et al., 2005).

In neither Tanzania nor Uganda has it been possible to track specific GBS effects down to the level of poverty-related outcomes or poverty impacts. The reasons include the relative infrequency of poverty measurement surveys, which means there are no data for periods that are relevant to plausible associations with the GBS inputs and their consequences. They also include real doubts about whether the expected mid-stream effects, other than volume of primary social service delivery, have actually (or have yet) been realised to the

degree that would be necessary for downstream improvements to be caused (ibid.).

2.3.3 *Impact evaluation*

The argument has been effectively made that impact evaluation deserves more careful attention in the development field. The evidence base for the development of new programmes in the social sectors is weak, and this is partly because systematic evaluation using best-practice techniques is too often not built into plans and budgets. What we do know from the rigorous impact evaluation work (e.g. that using randomisation techniques to handle otherwise difficult counterfactual issues) is that there can be substantial savings in scarce resources from using programme approaches that have been shown to have strong impacts and avoiding those whose impacts are questionable (Levine, n.d.; Levine et al., 2004; Savedoff et al., 2005).

According to White (2005) the term impact evaluation has several different meanings, ranging from rigorous analysis of the counterfactual to a focus on outcomes, evaluation carried out several years after an intervention has ended, and country wide or sector studies. Approaches to impact evaluation increasingly focus on randomised and/or experimental designs (Rawlings 2005). A key contribution of such techniques is the emphasis on paying greater attention to establishing controls in evaluation design. ***Such techniques are, nevertheless, only applicable in their most rigorous form to those types of intervention for which there is a clearly defined target group.*** Where this is not the case, for instance in the case of policy reforms, policy advice and institutional development, other approaches are needed.

White (2005) recommends merging a process-oriented approach with impact analysis. This may be practical for some purposes. However, it may be sufficient just to make the point that these are distinct but strongly complementary activities. We would suggest that certainly applies to the case of the Paris Declaration follow-up.

Savedoff et al. (2005) have made a good case that there is an Evaluation Gap at the impact end of the development evaluation results chain, especially where specific social-sector interventions are concerned. The practical relevance of these arguments to the DAC Evaluation Network is well treated in the World Bank IED's note on the subject (2006). It seems equally clear, however, that there is more than one field in which "what works" in development is less certain than it might be. ***Doing more to understand institutional and process issues is not an alternative to understanding impacts better, or vice versa. These are highly complementary agendas.*** This is especially so if we bear in mind that the Paris Declaration is centrally concerned with helping to create conditions in which there is a stronger constituency for results-based policy making and programme design.

3 Design of the evaluation work

3.1 Basic design issues

In this section we consider experience and perspectives relating to the three basic design areas covered in the Options Paper:

- ❑ the **scope** of the evaluation, especially the selection of units of evaluation and the range of themes to be covered;
- ❑ ways of ensuring **quality and consistency** in the evaluation work; and
- ❑ the **governance and management** arrangements appropriate to this type of exercise.

3.1.1 Scope of evaluation

A central question about the scope of a Paris Declaration evaluation effort is the degree to which it is feasible and wise to rely upon synthesising the results of country evaluations. White (2005) notes the value of country case studies when part of a broader “theory based approach”. He also argues for improvements in the way in which qualitative data are used to draw cross-country/generic lessons:

“It is rare to find evidence that qualitative data have been collected and analysed in a systematic way. For example, in a multi-country study, it is not sufficient to have common terms of reference. More detailed guidance is required on the range of data to be collected and the sources to be used. And these data need to be presented in such a way as to allow a systematic summary” (White, 2005).

White notes how the increased emphasis on results in the context of the MDGs has increased the importance of reports that aggregate across evaluations and the need for these reports to say something about outcomes. He also notes, however, that insufficient attention has been paid to methodological issues, in particular lack of explicit attention to the techniques of “meta-analysis” and the methodology used for aggregations. Drawing on Weiss (1998), he identifies six steps to meta-analysis:

- (i) define the problem,
- (ii) collect the relevant studies to be reviewed,
- (iii) screen the studies for quality,
- (iv) standardise indicators and code them,
- (v) produce a synthesis measure, and
- (vi) present results.

These steps apply specifically where there are quantitative, usually agency performance data, to hand. However meta-evaluations also require qualitative data. Combining qualitative and quantitative data points to the importance of a “theory based” approach (again following Weiss) that identifies core evaluation hypotheses which are subsequently tested using a mix of quantitative and

qualitative instruments and approaches. White refers to Rao and Woolcock (2003) for ways of integrating qualitative and quantitative approaches in programme evaluation.

The need to aggregate-up evidence of various types points to the need for detailed guidance on the range of data to be collected and the sources to be used in the country case studies. ***One of the things this makes clear is that synthesising the results of country case studies is unlikely to be a simple and inexpensive way of drawing global evaluation conclusions about the Paris approach.*** Nor is it at all likely to be compatible with the requirement not to impose additional burdens on partner countries.

3.1.2 *Ensuring quality and consistency*

A number of challenges have been identified in ensuring quality and consistency in joint evaluations (OECD, 2005):

- Developing comprehensive yet manageable terms of reference to accommodate a wide range of partner concerns and interests.
- The process of achieving a common understanding of the evaluation issues, since critical indicators and methods requires time and team interaction.
- Political sensitivities of all parties.
- Final agreement on report findings, which can become a major obstacle, including how divergent views are reported.

A OED review of experience with Joint Country Assistance Evaluations (World Bank, 2005) also notes the need to plan and budget separately for any evaluation capacity development support related to a joint evaluation, to avoid conflict of interest with the core business of the evaluation.

Picciotto and Feinstein (2005) note the close interaction between independence and quality. They argue that independence without competence can be misleading, costly and disruptive. Equally, quality is undermined if the evaluation process is characterised by a lack of criticality – objective assessment – and relevance to development effectiveness. Relevant skills, sound methods and transparency are required for any credible evaluation: “Thus independence and quality are synergistic” (3).

These observations confirm the importance of effective mechanisms for ensuring quality and consistency (alongside independence). Some of the ways of doing this include clear protocols relating to the way in which the evaluation process proceeds, guidance on appropriate and relevant sources of data, clear terms of reference regarding the purpose of the evaluation and flexibility in management of costs, especially additional time costs, related to a joint/multi-partner exercise.

The main mechanisms used to ensure quality and consistency in development evaluations include:

- A quality assurance technical advisor or group to oversee the preparation of TOR, communication of evaluative criteria, selection of evaluators and reviewing of evaluation reports.
- Engagement of evaluators – both nationals and donor staff – at country level to oversee use of evaluation criteria, independence etc.
- A common set of evaluation questions backed up by clear ground rules on data gathering, data sources and clear protocols on how data is to be aggregated to ensure consistent treatment across countries.

3.1.3 Governance and management

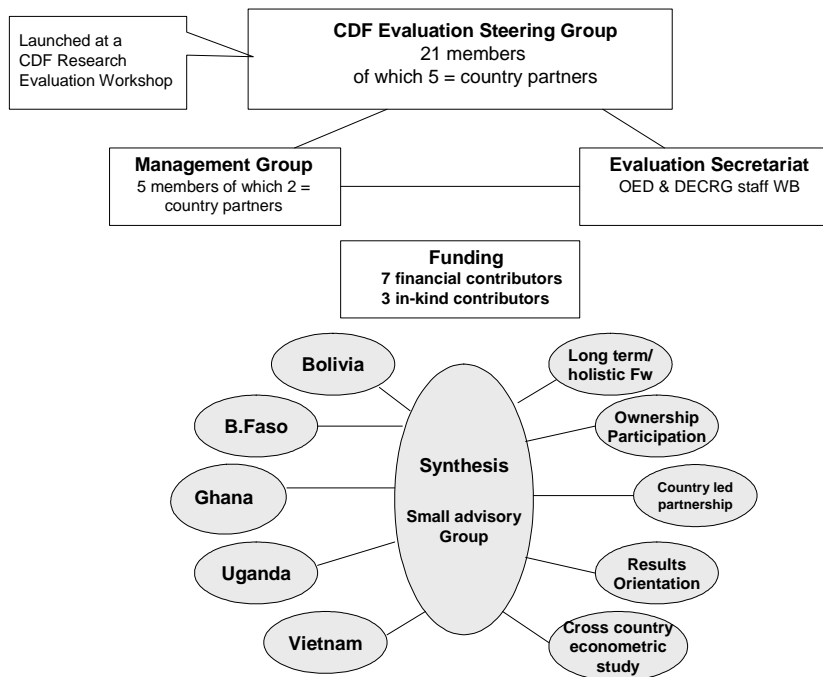
The new development “paradigms” – MDGs, PRSPs, aid effectiveness etc. – provide a strong case for joint evaluation work based on multi-partner approaches. The DAC Evaluation Network’s Working Paper on joint evaluations (OECD, 2005) cites various examples of multi-partner approaches. **Types of joint evaluation include:**

- 1) classic multi-partner – where participation is open to all stakeholders and partners participate on equal terms;
- 2) qualified multi-partner – where participation is open to those who qualify, e.g. as part of a specific donor grouping or as participants in a programme that is the subject of the evaluation; and
- 3) hybrid multi-partner – represented by a range of complex arrangements including delegated responsibility, parallel or inter-related evaluations, or joint activities focused around a common framework but implemented by devolved partners.

The review finds that there are few examples of 1). Most are either 2) or 3). The review notes that for the most part multi-partner evaluations are generated by individual actors negotiating through bilateral or group channels to solicit support and find partners. The review suggests it would be worth looking into the possibility of networking systematically between different groups (DAC, ECG, Nordic+ etc.) to a) find effective divisions of labour, and b) help in the development of common frameworks and ways to identify priority evaluation subjects of broad interest. **An evaluation of the Paris Declaration may be one such opportunity.**

One of the biggest challenges in multi-partner evaluations is arriving at a consensus on key findings and lessons. Much time is often spent at this stage, the more so the bigger the partnership. Ways of addressing the need to “gain consensus” vary. In the case of the meta-CDF evaluation, a central steering committee of some 21 members was formed to oversee the design and synthesis stages of the evaluation. This steering group was supported by a more streamlined management group and an evaluation secretariat at the World Bank. The management structure of the CDF evaluation is shown below.

Figure 3.1: Management structure of the CDF Evaluation



While the approach provided for fairly wide ranging country partner involvement at the start of the evaluation, the bulk of the work was in fact facilitated by the secretariat hosted by the World Bank and delivered by teams of Bank-based evaluators and northern-based consultants. ***In this sense it remained heavily “top down”.*** Funding was provided through the World Bank and a shared “pot” at the centre, with some in-kind contributions being made as part of the country studies. While this was effective and efficient in many ways, it is perhaps no surprise that the biggest stumbling block in the evaluation was gaining agreement at the “top” on the substantive evaluation findings and lessons.

Lessons from other multi-partner evaluations, including the recent GBS evaluation, confirm the need for early partner country engagement before the crucial design decisions are made. However, as the CDF evaluation experience shows, this does not rule out difficulties in trying to achieve a broad consensus on evaluation findings at the reporting stage.

An alternative approach is to avoid top-heavy management structures and to go for a more delegated model operating according to a fairly generic evaluation framework and core set of evaluation questions. Under this model each participating country would run the evaluation as it saw fit, while agreeing to report against a common set of questions and working to a very broad terms of reference. The ToR would only set out timelines, and lay down how to ensure independence etc. Funding and management responsibility would be completely devolved to the country level. Only the synthesis would be managed centrally and its actual preparation might be sub-contracted to an independent body.

3.2 Objectives and trade-offs

3.2.1 Stakeholder buy-in versus “criticality” of evaluation content

As early as 1991 the OECD DAC recognised the importance of stakeholder buy-in, stating:

“Development assistance is a cooperative partnership exercise between donors and recipients ... Both must therefore be interested in evaluation not only for improving resource use for development through learning from experience but also for accountability to political authorities and publics” (OECD, 1991).

For evaluations to be perceived to be useful and to be used they must feedback to stakeholders in a way that is relevant and appropriate to their needs.

At the same time, to be credible, evaluation must be seen as impartial and independent. Independent evaluation implies freedom from political influence, full access to information and full autonomy in carrying out investigations and reporting findings. The rise in the number of independent evaluation units in multilateral and bilateral agencies over the last ten years suggests that there is a strong demand for more independence in aid evaluation. Recently, there has been a call for more “third party” independent evaluation of the effectiveness of aid programmes and of aid agencies themselves (Savedoff et al., 2005; Levine, n.d.).

Independence must not, however, be confused with isolation (Picciotto and Feinstein, 2005). Ensuring that independent evaluation generates lessons that can be formative for those who design and manage aid programmes is crucial. Relevance is as important a criterion in doing evaluation as it is in judging the quality of an aid programme. Picciotto and Feinstein (2005: 3) note that “accurate and fair evaluations combine intellectual detachment with empathy and understanding. The ability to engage with diverse stakeholders and secure their trust while maintaining the integrity of the evaluation process is the acid test of evaluation professionalism”. ***Diminishing returns set in when evaluation independence assumes extreme forms.***

The way forward is therefore to ensure that evaluators can remain detached while interacting with programme managers, staff or beneficiaries – what is often termed *behavioural independence*. According to Picciotto and Feinstein (2005: 4) behavioural independence is a “privileged dimension of evaluation excellence” and can be secured through attention to four key areas:

- ❑ *criticality* – the ability and willingness to judge in an objective and transparent fashion;
- ❑ *additionality* – a distinctive contribution to development knowledge;
- ❑ *timeliness* – the delivery of operational evaluation findings in such a way as to inform decision-making and practice; and

- *materiality* – a focus on topics and issues that have direct relevance to understanding development effectiveness.

3.2.2 *Country learning versus cross-country learning*

There is no obvious way of reconciling the demand for evaluation work on the Paris commitments to contribute primarily to country-level learning, and the equally valid demand for insight into the global approach that the Declaration represents. That is so particularly if the global learning is conceived as requiring the synthesis of findings from country case studies, with all that that involves in terms of standardisation and rigour. However, that may be an excessively ambitious approach, particularly in view of what has been said about the GBS Evaluation and timescales of change.

In order for country evaluation work to assist country or regional learning processes, it needs to be allowed to vary in focus and approach from place to place. The most striking assertion of this perspective is the Hanoi Core Statement by the Vietnamese government and the Partnership Group on Aid Effectiveness in Vietnam, which “localises” the Paris Declaration in a way that reflects the particularities of the aid relationship in the country (PGAE, 2005).

Reports from the South-East Asian region suggest that understandings of what the Paris commitments mean are extremely varied, not only between different agencies and countries but among individuals as well (Beloe, 2005). As a joint initiative at the regional level, a number of country case studies of aid effectiveness initiatives are being commissioned. This effort is tailored to the practical conditions for taking forward Paris Declaration implementation in the region. There would be little advantage in imposing on it an international initiative to collect a large body of standardised information. On the other hand, it appears that there may be mutual benefit from a lighter form of engagement, in which representatives of the regional initiative participate in the discussion on a common evaluation framework and agree to contribute some data on selected core issues for international analysis.

Ways of doing this would have to differ from case to case. However, other regional events that have been held to discuss the implications of Paris are suggestive of similar possibilities (Entebbe, 2005). And countries such as Zambia and Nicaragua, which have invested quite heavily in monitorable Harmonisation and Alignment Action Plans (de Lange, 2005; Hammerschmidt, 2005) offer good prospects for country stakeholder buy-in to evaluation work, so long as the primary emphasis is on adding to what is already under way. A light engagement with a process to define a generic evaluation framework with a limited common core seems compatible with not imposing substantial new burdens on country officials, and could elicit real interest.

3.2.3 *Low costs versus value added*

Joint, multi-partner evaluations of global programmes of work are generally considered to be expensive, in terms of both direct and indirect costs. As regards direct costs, budgets for recent “global” evaluations have ranged between

€1million for the WFP evaluation and €2 million plus for the CDF and GBS evaluations (OECD, 2005).

While the direct cost to individual partners may not be greater than that of a stand-alone evaluation, it is often the case that the costs appear higher when they are not entirely under the control of one party. Indirect costs – staff time, the cost of travel to attend meetings and workshops and the hiring of services to cope with the increased workload of a joint evaluation – are inevitably higher. Part of the difficulty in managing the costs of joint evaluation lies in the very different human and budgetary capacities of evaluation departments, and the different level of “field presence” amongst partners supporting country-based evaluations.

Most of these costs are borne by agencies themselves, but costs relating to partner countries are less easy to enumerate given the diversity of different models of joint evaluation being used. Most direct costs are covered by donor agencies. Indirect costs include staff time directly involved in the evaluation, participation in steering groups, workshops etc. and the more indeterminate opportunity cost of engaging partner representatives in focus groups, interviews and requests for information. These costs may be significant. On the other hand, ***the costs need to be set against the benefits of a push towards more domestically-owned evaluations***, which imply (South Africa is a good example) stronger stakeholder participation (and local dissemination) and opportunities for evaluation capacity development.

From the perspective of donors the added value of a multi-partner approach comes from “muscle, backbone and legitimacy” (OECD, 2005). Joint evaluations can carry more weight, provide smaller donors with an important platform, tackle issues that affect the “community” of development partners and not just individual agencies, and build credibility with a larger group of relevant stakeholders.

Please refer to the accompanying Options Paper to see how these specific findings from the focussed literature survey have been translated into recommendations for Evaluation Network follow-up to the Paris Declaration.

References

- Balogun, P. (2005). *Evaluating Progress Towards Harmonisation*, Working Paper 15, London: DFID, May.
- Beloe, T. (2005) "Moving Towards Local Definitions of Aid Effectiveness: How is the OECD DAC Paris Declaration being Interpreted in DFID South East Asia Partner Countries?", Bangkok: DFID, Nov.
- de Lange, P. (2005) "Discussion Paper on the Relevance of a Joint Evaluation of Harmonisation in Zambia", The Hague: Ministry of Foreign Affairs, IOB, Mar.
- Elbadawi, I. and J. Randa (n.d.) "Assessing the Development Impact of CDF-like Experiences: Background Paper for A Multi Partner Evaluation of the CDF", Washington, DC: World Bank.
- Entebbe (2005) "Workshop on the Paris Declaration: Implications and Implementation, Entebbe, Uganda, Nov 16-17 2005", Workshop Report.
- Hammerschmidt, H. (2005) *Implementing Harmonisation & Alignment -- The Case of Nicaragua*, Report No. 3, Managua: SOFRECO, Apr.
- IDD and Associates (2006) "Joint Evaluation of General Budget Support 1994-2004: Draft Synthesis Report", Birmingham: University of Birmingham. IDD.
- Lawson, A and D. Booth (2004) *Evaluation of General Budget Support: Evaluation Framework*, London: DFID.
- Lawson, A. et al. (2005) *Does General Budget Support Work? Evidence from Tanzania*, London: ODI.
- Levine, David I. (n.d.) "Learning What Works – and What Doesn't: Building Learning into the Global Aid Industry", Berkeley: Haas School of Business, University of California.
- Levine, R. et al. (2004) *Millions Saved: Proven Successes in Global Health*, Washington, DC: Institute for International Economics.
- OECD (2005) *Joint Evaluations: Recent Experiences, Lessons Learned and Options for the Future*, Paris: DAC Evaluation Network, Working Paper.
- OECD (1991) *DAC Principles for Evaluation of Development Assistance*, Paris: OECD DAC.
- PGAE (2005) *Working Together to Improve Aid Effectiveness for Supporting Sustainable Development in Vietnam*, Hanoi: Partnership Group for Aid Effectiveness.

- Rao, V. and M. Woolcock (2003) "Integrating Qualitative and Quantitative Approaches in Program Evaluation" in F.J. Bourguignon and L. Pereira da Silva (eds.) *The Impact of Economic Policies on Poverty and Income Distribution*, New York: Oxford University Press.
- Rawlings, L. (2005) "Operational Reflections on Evaluating Development Programs" in G.K. Pitman (ed.) *Evaluating Development Effectiveness*. New Brunswick and London: Transaction Press.
- Picciotto, R. and O. Feinstein (2005) "Independence of the Operations Evaluation Department of the African Development Bank", Report of the Senior Evaluation Advisers, Washington, DC: World Bank, Oct.
- Rogerson (2005) "Aid Harmonisation and Alignment: Bridging the Gaps between Reality and the Paris Reform Agenda", *Development Policy Review*, Vol. 23, No. 5, Sept.
- Savedoff, William D.; Ruth Levine; and Nancy Birdsall (2005) "When Will We Ever Learn? Recommendations to Improve Social Development through Enhanced Impact Evaluation", Washington, DC: Center for Global Development.
- Weiss, Carol H. (1998) *Evaluation*, 2nd edn., New Jersey: Prentice Hall.
- White, H. (2005) *Challenges in Evaluating Development Effectiveness*, Working Paper 242, Brighton: IDS.
- World Bank (2006) "Impact Evaluation: An Overview and Some Issues for Discussion", Room Document 5 for 4th Meeting of the DAC Network on Development Evaluation, World Bank, Independent Evaluation Group.
- World Bank (2005) "Review of Experience with Joint Country Assistance Evaluations", Washington, DC: World Bank, Operations Evaluation Department.
- World Bank (1997-2005) Annual Review of Development Effectiveness, Washington, DC: World Bank, Operations Evaluation Department.