Evaluability Assessments: Reflections on a review of the literature

Rick Davies
Monitoring and Evaluation consultant, UK

Lina Payne
Department for International Development, UK

Abstract
In 2012 the United Kingdom’s Department for International Development (DFID) funded a review of the literature on Evaluability Assessments, which was undertaken by Rick Davies. Although the review focused on practical guidance Evaluability Assessments are not unproblematic. Evaluability Assessments involve an additional layer of cost and procedure. The diversity of evaluation approaches is a challenge to any categorical judgement about evaluability. The reach of an evaluability assessment can become over-extended. Evaluability questions asked of individual projects may not be so easily applied to larger portfolios of projects. The purpose of this article is to give more attention to the problematic aspects of Evaluability Assessments. In doing so it does not seek to revise the broad conclusions of the Working Paper, which unambiguously encouraged the wider use of Evaluability Assessments.

Keywords
DFID, evaluability, evaluation, guidance, literature review

The origins of the DFID Working Paper on Evaluability Assessment
Up until 2011, the UK Aid Department for International Development (DFID) jointly commissioned some large scale thematic evaluations, usually with other OECD DAC partners. These were informed by Approach Papers, which were the outcome of up to six months consultation and review, to focus evaluation questions on generating relevant and useful
evidence that would contribute to global evidence in a particular field. These Approach Papers would usually include some assessment of the existing evidence base, suggest evaluation questions and data availability and provide some comment on possible evaluation design; all of which would be further explored during an inception period for the ensuing evaluation. Since 2011, when DFID’s commissioning of evaluations shifted to a more decentralized model (primarily at country office level), very much smaller scale, up to five-day external Evaluability Assessments became the norm. Fourteen of these have been carried out by late 2014, representing approximately a quarter of their commissioned evaluations during this period.

Although DFID was now emphasizing the use of evaluations commissioned at country level, there were also central level commitments to assess the cumulative outcomes from activity representing new or renewed areas of investment across multiple countries. DFID’s policy commitments on support for gender equality, women and girls; work on empowerment and accountability; and DFID’s support to the International Climate Fund, were three such areas. Each evaluation would involve an assessment of a portfolio of projects. For each policy area, DFID commissioned a larger scale Evaluability Assessment, the aim of which was to clarify evaluation questions, draw boundaries around the scope of the evaluation, assess quality and availability of data, assess existing intervention logics, and develop an indicative budget and timeframe.

In November 2012 the DFID Evaluation Department commissioned a ‘Synthesis of Literature on Evaluability Assessments’. The impetus for this was recognition that Evaluability Assessments were being used in different ways within the organization, with very differing expectations, scope, process and timing. There was an assumption that Evaluability Assessments could represent good value for money, ensuring more focused, relevant and feasible commissioning of evaluations, along with a concern that DFID was not always getting the most from Evaluability Assessments, particularly in terms of their coverage. DFID wanted to learn more from the practice of other international partners, and improve their own practice going forwards. The need for clarity before commissioning evaluations became particularly important in a context where evaluations were being commissioned of highly complex programmes, often operating in fragile environments, and at a time when DFID (along with other development agencies) was demanding significantly more rigorous and credible evaluative findings.

The stated purpose of the literature review was ‘[t]o produce a short practical note that summarizes the literature on evaluability assessments, and highlights the main issues for consideration in commissioning an evaluability assessment’. The work was expected to ‘inform DFID and other development agencies about what an Evaluability Assessment can achieve, the scope of work, and issues to consider in commissioning Evaluability Assessments. The report should identify the main rationale for carrying out and using Evaluability Assessments, and identify current bottlenecks in their uptake and use’. The review was subsequently carried out by Rick Davies, the lead author of this reflection paper.

The literature on Evaluability Assessment

The review process began in November 2012. An English language document search via the internet generated a bibliography of 133 documents including journal articles, books, reports and web pages, covering the period 1979 to 2012. Of these, 59 percent described actual examples of Evaluability Assessments, 13 percent reviewed experiences of multiple kinds of Evaluability Assessments, 28 percent were expositions on Evaluability Assessments, with
some references to examples, 10 percent were official guidance documents on how to do Evaluability Assessments and 12 percent were Terms of Reference for Evaluability Assessments. Almost half (44%) of the documents were produced by international development agencies. The majority of the remaining documents were produced by state and national agencies in the USA. Documents on international practice were largely by non-American agencies (i.e. 88% of the documents reviewed). The bibliography is now available online, and includes additional sources identified since 2012.2

Figure 1 below illustrates changing trends in the use of Evaluability Assessments, since the 1970s,3 as reflected by number of publications per year. As has been noted by previous reviews, Evaluability Assessments became popular in the USA in the 1980s, after they were promoted by Joseph Wholey (1979), an American government official. They subsequently fell in to disuse for a decade but have since seen a resurgence in use in the last decade. It has only been in the last 15 years that international development agencies have begun to use Evaluability Assessments. This interest has been especially notable in very recent years. As noted in the DFID Working Paper, in the previous year guidance on Evaluability Assessment had been developed by ILO, CDA, IDRC, ERBRD and UNODC.4 At the time of the Working Paper, 12 Evaluability Assessments had been carried out by seven different international development agencies in the past year.5

Very few of the Evaluability Assessments carried out by international development agencies made any reference to prior experiences with Evaluability Assessments. A few made reference to the more widely cited American commentators in the field, such as Wholey, Thurston, Smith, Leviton and Trevisan. Similarly, there were negligible references to international experiences with Evaluability Assessment in the American domestic literature. This has
since changed with a recent book on Evaluability Assessment by Trevisan and Walser (2014: 8) noting ‘the vibrant Evaluability Assessment work associated with international development evaluation’. In turn, this reflection paper has made use of Trevisan and Walser’s summary of American domestic experience.

**The Working Paper’s analysis and recommendations**

The results of the literature review were published as a DFID Working Paper in 2013 (Davies, 2013). The paper covered four broad areas: (a) the purpose of an Evaluability Assessment; (b) planning for an Evaluability Assessment; (c) the process of doing an Evaluability Assessment; and (d) potential concerns about Evaluability Assessments. The Working Paper included 18 recommendations and provided draft Terms of Reference for an Evaluability Assessment along with an evaluability checklist. Central to the guidance is a very specific definition of Evaluability Assessment, its expected outputs and the checklist, which provides a menu of specific questions to ask during an assessment and covers three separate dimensions of evaluability.

The Working Paper accepted the OECD DAC (2010: 21) definition of evaluability as ‘[t]he extent to which an activity or project can be evaluated in a reliable and credible fashion’. Evaluability assessment was then defined in the Working Paper as having three related dimensions:

- Evaluability ‘in principle’, as seen in the quality of the project design;
- Evaluability ‘in practice’ as seen in the availability of data;
- The utility and practicality of an evaluation, as seen in the institutional context.

The DFID Working Paper was designed to provide straightforward practically oriented advice. However, behind many of the recommendations are some problematic issues to do with how Evaluability Assessments can be carried out. Some of these are evident in a reading of the largely American domestic literature on Evaluability Assessments. Other issues are emerging as DFID seeks to carry out Evaluability Assessments of large portfolios of projects spanning multiple countries, as distinct from evaluations of specific projects. The Working Paper took a position on some of these issues but did not address others. The purpose of this reflection paper is to explore these issues in more detail, and to propose some possible ways forward. In doing so it does not seek to revise the broad conclusions and recommendations of the Working Paper, which unambiguously encouraged the wider use of Evaluability Assessments.

**When is an Evaluability Assessment not needed?**

In contrast to the benefits expected from an Evaluability Assessment, there has been negligible discussion in both the domestic and international development agency literature on when their use is not appropriate. This issue was also not covered by the Working Paper. This is ironic given that one of the original and continuing purposes of Evaluability Assessments has been to decide when an evaluation may or may not be appropriate.

On reflection, two broad circumstances can be identified and need to be highlighted. First, there will be instances where a project design is manifestly unevaluable. In one project seen by one of the authors in the last year, both the Theory of Change and the monitoring and
evaluation provisions were unusable and this was recognized by the project manager. The most appropriate next step was to develop an M&E framework that addressed these issues, not to commission an Evaluability Assessment. The second circumstance relates to questionable cost-effectiveness. In small projects, such as those implemented by some NGOs, with even smaller evaluation budgets an Evaluability Assessment will need to be very cheap indeed before it could be expected to make a cost-effective difference to the effectiveness of the evaluation.

Further development of guidance on Evaluability Assessments could usefully emphasize that Evaluability Assessments are not compulsory or necessary in all circumstances. If they were promoted or perceived as a requirement then this could well be the kiss of death for Evaluability Assessments, leading to their ritualization and then eventual neglect. It is worth noting that even the Inter-American Development Bank (IADB), which has carried out more Evaluability Assessment than most international development agencies, has only carried them out on a random sample of their projects.

**What are the alternatives to an Evaluability Assessment?**

The issue of alternatives to Evaluability Assessments has not been discussed in the international development agency literature on Evaluability Assessment and was not covered by the Working Paper. However, a number of options have been identified. One used by the IADB is to expand the ambit of the quality assurance process used when approving newly designed projects for funding, to include criteria relating to evaluability. This was done after a decade’s experience with the use of separate Evaluability Assessment on an annual sample of project proposals. Another is the use of independent procedural audits, to examine the functioning of project monitoring systems, as are now being supported by the Donor Committee on Enterprise Development (2014). The DCED audit process is comprehensive and now well tried, covering an even greater number of areas than most Evaluability Assessments. UNAIDS (2014) has developed a similar tool for three areas of its work. Like the Working Paper, the core element is a set of checklists. Unlike the DCED procedural audits, it is designed as a means of self-assessment.

When faced with large scale evaluations involving multiple agencies and across a number of country programmes, donors such as DFID have previously supported the use of Approach Papers, preceding the commissioning of evaluations. These seem to cover similar territory to Evaluability Assessments but differ in that they draw significantly on literature reviews to identify gaps in the global evidence for a particular thematic area; they also tended to involve a more systematic process for engaging with external stakeholders. It could be argued that while they placed more emphasis on utility, relevance of evaluation questions and evaluation design options, they gave insufficient attention to programme theory (i.e. evaluability in principle) or an examination of the data availability (i.e. evaluability in practice). The Approach Paper for the Multi-stakeholder Evaluation Public Sector Governance Reform (PSGR) in 2010, for example, included text on the relationship to previous research and learning, as well as a suggested methodological approach, but there was a limited analysis of the datasets available. If Approach Papers are to be used in the future, it would be important that they addressed all three dimensions of evaluability.

In recent years, where DFID evaluations have shifted to focus more on single project and programme level assessments, Inception Phases have sometimes been used to carry out Evaluability Assessments (although these phases have not been required to follow any systematic process for evaluability). In the author’s own recent experience this practice has not
been without problems. In a 2013 evaluation of DFID’s investments in rural livelihood projects in India, the Inception Report noted many problems with absent data and non-existent portfolio level Theories of Change. Nevertheless, the evaluation did go ahead but was terminated in mid-process some time later. It could be argued that a separate Evaluability Assessment by a third party may have led to an earlier conclusion not to proceed. There are inbuilt incentives to proceed where evaluability is contractually built into an ensuing evaluation. Evaluations of other DFID funded programmes such as Girl Hub seem to be proceeding on the same basis, with Evaluability Assessments being built in as the first part of a single evaluation contract, and may well face the same risks. As noted in the Working Paper, USAID experience suggests that splitting the work into separate contracts, even for the same contractor, may be the better option, because it builds in the serious option of not proceeding with an evaluation.

Another alternative to Evaluability Assessments is the use of pilot testing of specific evaluation methods. Pilot testing is potentially useful when a particular design is being proposed but the parameters of that methodology need clarification. Here the resulting decision is more likely to focus on how to proceed rather than whether to proceed.

**How ambitious should an Evaluability Assessment be?**

The core purpose of an Evaluability Assessment is to assess evaluability, which has been defined by the OECD DAC as ‘[t]he extent to which an activity or project can be evaluated in a reliable and credible fashion’. However, the literature on Evaluability Assessment has identified a number of wider objectives (five by Leviton et al., 2010: 1) and benefits (17 by Trevisan and Walser, 2014: 139).

In actual practice, the ambitions of an Evaluability Assessment can segue into other types of evaluation practices. This tendency seems to be almost inbuilt into Wholey’s (1979) original conception of Evaluability Assessment, when he describes it as ‘exploratory evaluation’, a view he has repeated in his most recent publications (2010). Trevisan’s (2007: 300, 288) review of 20 years practice found evidence of Evaluability Assessments morphing in various directions:

Although Evaluability Assessment continues to be used for its original purpose of determining the readiness of a program for impact assessment, 17 of 23 articles presented Evaluability Assessment studies that were at least in part designed to perform a formative evaluation function. Further obfuscating the Evaluability Assessment concept is the use of formative evaluation, process evaluation, and needs assessment, in some studies.

He concluded that ‘clear differentiation between Evaluability Assessment and other evaluation processes (such as process evaluation and needs assessment) is recommended’.

As identified by the Working Paper, the objectives of an Evaluability Assessment can also vary according to its timing. When carried out at an early stage in a project cycle, an Evaluability Assessment can have useful implications for project design, as represented in a Theory of Change and elsewhere. The question then is how far to go, beyond the identification of design problems? In practice many if not most clients will want to see some resolution of these problems, in the form of a more workable Theory of Change at the least. At a later stage in the project cycle the analysis of data availability will have implications for the design of M&E plans and procedures. To what extent should these be resolved versus simply identified? When carried out prior to an evaluation it will be easy for the Evaluability Assessment to segue into
evaluation design. How far should the Evaluability Assessment proceed into that territory? There does not seem to be a clear basis on which to identify where the boundaries should be drawn. In practice it may depend on the client’s institutional funding and approval mechanisms and the resources available. Most of these extensions of purpose will require significantly additional stakeholder involvement.

The Working Paper took a conservative position on the issue of breadth of purpose, arguing that an Evaluability Assessment should present a diagnosis, which could then form part of the Terms of Reference for an Evaluation, or the Terms of Reference for the design of an M&E framework at an earlier stage in the project cycle, which would then address the issues raised.

What kind of guidance is needed?

Are stage models useful? The Working Paper identified 12 different stage models, proposed in the international development agency and American domestic literature on Evaluability Assessments, and this list is by no means exhaustive. Despite this apparent wealth of choice, Trevisan’s (2007) review of 20 years use of Evaluability Assessments noted that ‘[f]or those studies that did cite known Evaluability Assessment authors, most did not indicate that the steps offered by a particular author or authors were specifically followed or that the findings logically followed from these steps’.

More recently, the use of stage models has been criticized for their linearity. Leviton (2010: 220) used a flow chart to allow for re-iteration in the assessment process and Trevisan and Walser (2014) used a circle with four quadrants to illustrate the iterative relationship between what they emphasize are components of an assessment rather than steps. Their model has then been supported by checklists of questions for each of the four components. This is in positive contrast to a number of earlier stage models which had unhelpfully mixed up both sequences of activities and checklists of activities (Davies, 2013: 43–4).

The Working Paper proposed a relatively simple and self-evident sequence of stages, with recognition that there would be some re-iteration between steps 2 to 4, as shown in Table 1. This sequence was supported by the use of checklists, but these were not specific to each stage. This was because many of the questions in the checklists could be asked of documents, people or both.

Are checklists of any value? The 2012 literature review suggested that the use of checklists seems to be much more common amongst international development agency practice than it is amongst American domestic agencies. More than half (11) of the 19 international agencies found to be using Evaluability Assessments have used checklists in one form or another. Their proposed uses of checklists varied substantially, from quasi-mandatory to highly optional,

Table 1. Stages in an Evaluability Assessment.

<table>
<thead>
<tr>
<th>Stage</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Define the boundaries of the project to be assessed and the scope of the Evaluability Assessment</td>
</tr>
<tr>
<td>2.</td>
<td>Identify the documents and stakeholders accessible to the Evaluability Assessment</td>
</tr>
<tr>
<td>3.</td>
<td>Review documents Using checklists of issues on three dimensions of evaluability</td>
</tr>
<tr>
<td>4.</td>
<td>Engage with stakeholders</td>
</tr>
<tr>
<td>5.</td>
<td>Develop conclusions and make recommendations</td>
</tr>
<tr>
<td>6.</td>
<td>Feedback findings and conclusions to stakeholders</td>
</tr>
</tbody>
</table>
from very structured response options to the more openly formatted. There were many ways of using checklists. This diversity in use probably reflects varying views on the merits and status of checklists.

However, the Working Paper recommendations clearly supported their use for a number of reasons: (a) checklists are a means of ensuring the systematic and accountable examination of all relevant issues; (b) checklists can be structured into sections and sub-sections, enabling meso-level judgements to be built up from micro-judgements; and (c) the design of checklists can be informed by relevant theory. The latter point has been emphasized by Michael Scriven (2007: 1) in his defence of checklists:

The humble checklist, while no one would deny its utility in evaluation and elsewhere, is usually thought to fall somewhat below the entry level of what we call a methodology, let alone a theory. But many checklists used in evaluation incorporate a quite complex theory, or at least a set of assumptions, which we are well advised to uncover; and the process of validating an evaluative checklist is a task calling for considerable sophistication. Indeed, while the theory underlying a checklist is less ambitious than the kind that we normally call a program theory, it is often all the theory we need for an evaluation.

The Working Paper reviewed the checklists that have been used by international development agencies and synthesized their contents into three separate checklists reflecting what were seen as the core dimensions of evaluability (Davies, 2013: 20–3). These are in effect a mini-theory of what constitutes evaluability. Their construction was not explicitly informed by reference to wider theories, but on reflection some relationships can be identified which could help inform future revisions.

The first is the relationship with US Joint Committee Program Evaluation Standards (Yarbrough et al., 2010), as shown in Table 2 above. These were used to inform the contents of checklists used in the Evaluability Assessment model recently developed by Trevisan and Walser (2014: 161).

The second set of relationships is with the categories of evaluation design choices described in Stern et al. (2012) in their report on ‘Broadening the Range of Designs and Methods For Impact Evaluations’. They point out that the selection of appropriate evaluation designs has to satisfy three kinds of constraints or demands. These demands and how they are informed by aspects of evaluability are listed in Table 3 above.
There is a third body of theory and evidence that the Working Paper checklists did not attend to but which could also inform future revisions. This is the extensive research literature on evaluation use (e.g., Johnson et al., 2005).

How can the findings be synthesized? Trevisan and Walser’s (2014) recent book on Evaluability Assessment provides the most comprehensive coverage of evaluability issues since the last book was published on the topic 25 years ago. However, their coverage of the analysis of evaluability data is brief relative to the diversity of data collection methods which they suggest can be used for Evaluability Assessment (Trevisan and Walser, 2014: 118–19, 133–5). This is unfortunate. In the authors’ own experience the time required for data analysis is often underestimated in evaluation planning and the process of analysis, especially of qualitative data, is often not transparent. DFID have recently produced an internal ‘How to Note’ on the analysis of qualitative data, because of the same concerns emerging from both evaluation and research commissioned. The development of further guidance on the synthesis of Evaluability Assessment findings could help address this problem.

The Working Paper did address three aspects of the synthesis task, under headings relating to the calculation of aggregate scores, the use of weightings and the question of whether evaluability can really be measured. The calculation of scores on different aspects of evaluability is useful where multiple projects are being screened as part of one Evaluability Assessment exercise. This was the case with the Evaluability Assessment of SIDA’s human rights projects (SIDA, 2000). The authors were able to identify the relative prevalence of different evaluability issues across the portfolio. The Inter-American Development bank (IADB) has used evaluability scores to make systematic comparisons of project designs across different evaluability dimensions and across different project types, on three occasions since 2001. The findings influenced the IADB’s decision to change their project design quality assurance process.

Most international development agency evaluability checklists do not, however, use formal weightings; but any attempt to aggregate scores will implicitly, if not explicitly, assign weights to component scores. Some agencies like the International Labour Office (ILO) have pre-defined weights built into their instrument, which presumably apply to all of their Evaluability Assessments. The Working Paper proposed a more flexible alternative, which is that weights should be assigned by those doing a particular Evaluability Assessment. This would enable the particular intentions of specific evaluations to be taken into account. For example, if there were plans for a process evaluation versus an impact evaluation.

The question of how to decide when a project is or is not evaluable, given all the data that have been collected, including any aggregate and weighted scores, is inherently problematic. Reflecting on their experience of Evaluability Assessment in Burkina Faso, D’Ostie-Racine et al. (2013: 78) reported that ‘[s]eeking through the literature to see how other researchers have dealt with this issue proved to be futile because to our knowledge no authors have addressed this point’. This is not surprising given the proliferation of evaluation methods and approaches. For example, a goal-free evaluation would not be too worried about the absence of an evaluable Theory of Change, a point argued by an International Development Research Centre (IDRC) paper on Evaluability Assessment (Monk, 2012). The Working Paper’s response was that any aggregate scoring of evaluability (or its components) should be treated as an ‘index of difficulty’, which would signal to evaluators and their clients where the challenges were that most needed to be attended to in evaluation proposals. In the absence of the use of evaluability scoring, it should be possible for an Evaluability Assessment to identify, in
a given context, any problematic conditions that constitute any obstacles that must be dealt with before an evaluation can proceed.

Ideally, the Working Paper’s guidance could be further developed by separating three levels of the synthesis task. At the base is the synthesis of answers from multiple documents and interviews in respect to a specific checklist question. Here the assessment needs to be about validity and reliability of the data and then the identification of what the consensus and outlier views are. At the next level is the synthesis of answers across multiple questions within a given evaluability dimension. Here the assessment needs to ask about the existence of any ‘obstacle’ problems and then amongst the others, a weighting of their relative importance. At the next level, looking across the three dimensions of evaluability, the assessment needs to be about how the answers under each dimension fit with each other.

• How do the available data fit the Theory of Change?
• How do the stakeholders questions relate to the Theory of Change?
• How do the available data relate to the stakeholders questions?

An Evaluability Assessment’s answers to these questions will involve the identification of problems, where they exist, and suggestions or recommendations as to how they can be resolved. How much attention is given to the latter will depend on the ambit of the Evaluability Assessment, an open issue that has been discussed above. In the case of the Evaluability Assessment of DFID’s strategic vision for women and girls, and empowerment and accountability policy areas, whilst an exclusive focus on checklist questions would have determined that an evaluation was not possible, a wider perspective on the findings identified a constructive way forward, which included laying the groundwork for improved data accessibility.

The particular challenges faced by international development programmes

The Working Paper made the reasonable assumption that the focus of most Evaluability Assessments would be individual programmes, or at least a set of programmes with a common purpose and strategy. This is likely to be the case when an Evaluability Assessment is carried out within a single country. However, international development agencies often fund development programmes across a range of countries, which address a particular policy area of interest. In recent years within DFID there have been at least five planned evaluations of portfolios of projects spanning multiple countries: DFID’s policy approach to empowerment and accountability (E&A), DFID’s strategic vision for women and girls (SVW&G), the operations of the International Climate Fund (ICF), DFID’s approach to Security and Justice, and the Africa Regional Department’s Programme for Empowerment and Accountability in Africa. Evaluability Assessments were commissioned for each of these portfolios. These assessments have subsequently informed the design of DFID’s Terms of Reference for the evaluations. In the first two cases the Evaluability Assessment recommended that an evaluation was not feasible until issues to do with data availability and Theory of Change were resolved; very similar issues emerged from the security and justice and ICF related work. The ICF Evaluability Assessment made a series of recommendations concerning building a clearer monitoring, evaluation and learning strategy, incorporating an overarching Theory of Change,
improved mapping and documentation via a dedicated document archive, and developing a clearer quality framework. Although DFID recognized the results challenged their preconceptions of what was possible, they nevertheless viewed the Evaluability Assessment process as valuable. The Evaluability Assessments were seen as delivering good value for money, representing a small fraction of the likely cost of an evaluation, yet were subsequently assumed to be having a significant influence on the design and direction of those evaluations.¹¹

Evaluations of multi-country programmes like those above of DFID face a number of common challenges, arising from their core characteristics of greater diversity and more limited central control, compared with the evaluation of within-country programmes. These challenges relate to the three areas that are addressed by the Evaluability Assessment framework discussed above:

- Very complex yet incomplete Theories of Change;
- Diverse, incomplete and unknown data;
- Numerous and scattered stakeholders.

**Very complex Theories of Change**

In all five of the DFID evaluations, the intervention logic or programme theory was necessarily extremely broad, representing multiple change processes at a macro level and often via unspecified mechanisms. While each portfolio has a generic high level outcome the more specific and measurable forms of their outcomes often varied from country to country. In addition, the fact that interventions are taking place in multiple countries means that the intervention contexts are very diverse.

The Working Paper’s expectation that an evaluable project design will have identifiable, and thus testable causal pathways, was not realistic at this level. Any attempt to represent all causal pathways that connect contexts, interventions and outcomes in a multi-country portfolio of projects is likely to look like a bowl of spaghetti.

Despite these problems, the portfolio-level Theories of Change for the kind of programmes described above can contain two useful kinds of lists.¹² These are the outcomes expected and the kinds of interventions planned. These represent accountable results and resource commitments respectively and remain minimal requirements for some form of accountability. If these are intact, or can be reconstructed, then there is the possibility that an evaluation can develop a post-hoc construction of an evidence-based Theory of Change of what actually happened.¹³ Even prior to an evaluation, as part of an Evaluability Assessment, hypotheses can be sought from stakeholders about which specific outputs are expected to causally contribute to which specific outcome levels changes and in what circumstances.

Before this can be done, the problem of lack of shared outcomes needs to be resolved. In the case of the first two DFID portfolios above the proposed solution was to dis-aggregate the portfolio into identifiable clusters of projects which did have some comparable measurable outcomes. That process is now underway following the commissioning of two macro-evaluations. Other possibilities have also been identified, including the identification of clusters of projects based on a common intervention. If these can be identified then an ‘effects of a cause’ analysis could be feasible, in contrast to a ‘causes of an effect’ analysis that may be possible with an outcome centred cluster of projects.¹⁴
Diverse, incomplete and unknown data

The simplified stage model for an Evaluability Assessment, discussed above, proposed an initial step whereby boundaries were identified as to what would be included in an evaluation. In the Evaluability Assessment of two DFID portfolios this initial task was problematic. There was a lack of information on what constituted memberships of the portfolio, and what lay outside. While there is some systematic coding of DFID project records in relation to gender issues, it did not cover the range of projects the objectives addressed by the Strategic Vision for Girls and Women (SVG&W). There was no comparable coding of empowerment and accountability projects. Evaluations of other kinds of portfolios may face the same boundary definition problems, unless they have their own dedicated funding window(s), such as the ICF. Project documentation which would enable this kind of coding at a later date was also largely inaccessible, despite the recent development of DFID’s Development Tracker online project database. The database, which does have great potential, is designed to include a wide range of documents including project designs, financial data and progress reports. Access to these documents is essential if any attempt is to be made to code and cluster projects by type of context, intervention or outcome.

At the time the Evaluability Assessments were carried out there were still significant gaps in the database’s coverage of evaluations and related materials such as annual and mid-term reviews. This was significant because the original proposed design of the portfolio evaluations focused on the use of existing and planned evaluations of component projects, alongside a quality assurance process that would ensure component evaluations were of good enough quality. To do this would require not only data on evaluations already carried out, but also on those planned, which in practice is more challenging. Development Tracker does not include that level of detail and it would be questionable whether data on planned evaluations could ever be very complete or reliable.

In the case of both the empowerment & accountability and strategic vision for women and girls portfolios, the net consequences of these data availability problems were that a portfolio-evaluation was at risk of evaluating a sample of projects from an unknown population, using a sample of data from an unknown population of data sources. The problem of inadequate data was seen as more significant and less tractable than the problems relating to complex and incomplete Theory of Change. The suggested solutions to the Theory of Change problem were reliant on relevant data being available. In their existing state, they were judged to be unevaluable. By highlighting this significant constraint, the Evaluability Assessment enabled a way forward to be identified, in this case a TOR for the evaluation that included a lengthy project mapping, tabulation and documentation process.

Numerous and scattered stakeholders

The identification of stakeholders in a portfolio evaluation in international development is arguably an order of magnitude more difficult than the identification of available data. Stakeholders associated with projects are scattered across numerous countries and diverse institutions. They can include global audiences interested in the policy area being evaluated. Databases like the Development Tracker do not record data on stakeholders as a specific searchable category of information. At best the names of staff or individuals associated with projects may be referred to in documents such as project Annual Reviews.
Another constraint is resources available. Because budgets for Evaluability Assessments need to be kept in proportion to evaluation costs it may not be practical to conduct extensive consultations with country level stakeholders. Evaluability Assessments that have been carried out of DFID’s portfolios of projects have often focused on a relatively small circle of stakeholders readily identifiable by those commissioning the Evaluability Assessments. The ICF Evaluability Assessment interviewed 15 stakeholders, the E&A and SVW&G Evaluability Assessments interviewed eight and 12 stakeholders. The Evaluability Assessment of Security and Justice portfolio was notable for its wider coverage, of 30 stakeholders, including 20 from countries hosting the projects of interest. Where consultations have been limited, this has meant that the examination of the relevance of evaluation questions has in effect been skewed, towards the interests of those who are available and likely to be interested.

The net effect of these sources of constraint is that the third dimension of evaluability, which is concerned with practicality and utility, has been the weakest. Questions could then be asked as to whether more attention to this aspect of evaluability could help resolve some of the problems with data availability. Willingness of project managers to tune aspects of their own forthcoming evaluations to the needs of a portfolio evaluation might be increased – as might the supply of crucial but often missing project documentation. If the interests of project level stakeholders are more local than those apparently being addressed by a portfolio evaluation, then an array of other measures may also be needed in addition to simply investing more time. Clustering of projects by common outcome, as proposed above, may increase the salience of a portfolio evaluation. Canvassing of relevant evaluation questions using short structured online surveys may help widen understanding of a wider range of project level stakeholders’ interests. But these steps alone are not likely to be sufficient.

**Some interim conclusions**

Guidance documents are just that. It is likely they will continue to be re-interpreted and probably re-combined with other documents in new ‘mashups’, according to institutional needs. As pointed out above, the 2012 Working Paper could already be updated in the light of recent publications on American domestic experience (Trevisan and Walser, 2014), and DFID’s experiences with policy area evaluations. Future guidance should incorporate improved advice on:

- The circumstances when an Evaluability Assessment may not be needed;
- The alternatives to doing an Evaluability Assessment;
- The importance of clearly bounded expectations of the outputs of an Evaluability Assessment;
- The value of linking checklist contents to relevant wider theory and evidence;
- The need for an explicit process for analysing and synthesizing results of an Evaluability Assessment.

The Working Paper has already noted that Evaluability Assessments have been carried out in a range of areas, not simply single project level evaluations. As the range of their use expands there will be new questions arising as to how to best carry out such Evaluability Assessments. This is especially evident in DFID’s recent commissioning of Evaluability
Assessments of policy areas (i.e. portfolios of projects covering a common policy objective, but which are often scattered across the globe). In this article three broad problems were identified, which relate to the three dimensions of evaluability:

- Very complex yet incomplete Theories of Change
- Diverse, incomplete and unknown data
- Numerous and scattered stakeholders
- Evaluability in principle
- Evaluability in Practice
- Institutional context (practicality and utility)

The review of DFID’s experience with a small number of portfolio level evaluations suggests that problems with evaluability of project designs with complex and incomplete Theories of Change are tractable, if problems with data availability are manageable. But problems with data availability are significant and they seem to have at least some of their roots in the difficulties of engaging with distant and diverse stakeholders. These difficulties in turn are related to project level stakeholders having relatively autonomous relationship with the more centrally located commissioners of portfolio evaluations. The challenge for both Evaluability Assessments and evaluations themselves is how best to reach out and effectively engage with those stakeholders, within the constraints of time and resources likely to be available to them. This is an issue that further versions of guidance on Evaluability Assessment should seek to clarify and make use of any examples of emerging good practice.

Notes

1. The Evaluability Assessment bibliography was compiled on the basis of:
   - Searches via Google Scholar and Google Search to find documents with ‘evaluability’ in the title. The first 100 items in the search result listing were examined.
   - Searches via PubMed, JSTOR and Sciverse using the same keyword, and with the same limit within each search result.
   - An inquiry made via the MandE NEWS, Xceval and Theory Based Evaluation email lists.
   - Scanning of references within academic documents on evaluability found with high citation counts and within Evaluability Assessments and guidelines produced by international development agencies.
   - References referred to by interviewees within international development agencies.


3. Bear in mind this chart may also reflect the greater accessibility of the most recent documents, relative to older (pre-internet) documents. Leviton et al. (2010) reports that ‘[i]n the late 1970s and 1980s, more than 50 evaluability assessments were conducted, 33 of these between 1980 and 1985. Wholey left government and use of the technique dropped off significantly. Between 1986 and 2000, only eight evaluability assessments could be identified’. The search results may also under report the amount of Evaluability Assessment work being undertaken within the USA in the last decade. Leviton (2010) has identified 50 in the field of public health.

4. International Labour Office (ILO), Centre for Development Action (CDA), International Development Research Centre (IDRC), European Bank for Reconstruction and Development (ERBRD) and United Nations Office on Drugs and Crime (UNODC).

5. DFID, AusAID, United Nations Family Planning Agency (UNFPA), World Food Program (WFP), Inter-American Development Bank (IADB), United Nations Development Fund for Women (UNIFEM) and HELP (a German NGO).

6. Defined as ‘the process of change by outlining causal linkages in an initiative, i.e., its shorter-term, intermediate, and longer-term outcomes. The identified changes are mapped – as the “outcomes
“pathway” – showing each outcome in logical relationship to all the others, as well as chronologica
flow. The links between outcomes are explained by “rationales” or statements of why one outcome
is thought to be a prerequisite for another’ (http://en.wikipedia.org/wiki/Theory_of_change).
7. See Trevisan and Walser (2014: 3–9) for references to other models by other authors.
8. Published by Wholey et al. (1979).
9. Approximately 60% of the Evaluability Assessments reviewed in the Working Paper were of indi
vidual projects (Davies, 2013: 10). The rest were Evaluability Assessments of country strategies,
strategic plans, workplans, partnerships and sets of projects within particular policy areas.
10. In the case of the ICF a mid-term evaluation was already underway, but the assessment is informing
the TOR for a wider Monitoring Evaluation and Learning contract.
11. The cost of the Evaluability Assessment of empowerment and accountability and strategic vision
for women and girls portfolios was estimated at 5% of the project cost of the proposed evalua
tion. The cost of the Evaluability Assessment of the ICF represented 1.3% of the value of the por
tfolio.
12. For example, the ICF has a list of Key Performance Indicators that refer to events at the Output and
outcome level.
13. For example, using theory driven method like Qualitative Comparative Analysis (QCA) or data
mining algorithms that can identify clusters and association rules.
14. The possibility of this kind of analysis has been elaborated here: http://mandenews.blogspot.
couk/2013_07_01_archive.html.
15. Available at: http://devtracker.dfid.gov.uk/location/country/ [accessed 16 September 2014].

References
based Non-Governmental Organization’s (NGO) progressive evaluation strategy. Evaluation and
Gilchrist J (2014) Evaluability assessment: developing a process to determine if a program is ready and
able to be evaluated. Thesis. Simon Fraser University. URL (consulted 30 July 2014): summit.sfu.
Hare J and Guetterman T (2014) Evaluability assessment: clarifying organizational support and data
OECD (2010) Glossary of key terms in evaluation and results based management (in English, French
keepeek.com/Digital-Asset-Management/oecd/development/evaluation-and-aid-effectiveness-
no-6-glossary-of-key-terms-in-evaluation-and-results-based-management-en-english-french-and-
spanish_9789264034921-en-fr#page2.
Scriven M (2007) The logic and methodology of checklists. Western Michigan University. URL (consulted
8 September 2014): http://www.wmich.edu/evalctr/archive_checklists/papers/logic&methodology_
dec07.pdf.
Stern L, Stame N, Mayne J, et al. (2012) Broadening the range of designs and methods for impact evalua
tions: report of a study commissioned by the Department for International Development. Working


Rick Davies is an independent evaluation consultant, working with international development aid organisations.

Lina Payne is an Evaluation and Social Development Adviser in Evaluation Department of the Department for International Development (Note this was her position at the time the paper was written. It has since changed. She is now the Social Development Adviser in the Research & Evidence Division of the Department for International Development).