Evaluability Assessment

For DFID's
Empowerment and Accountability
And Gender Teams

Report
26 July 2012

Rick Davies
Sarah-Jane Marriott
Sam Gibson
Emma Haegeman

theIDLgroup

Church & Court Barn, Church Lane
Tickenham, Bristol, BS21 6SD
United Kingdom

Tel: +44 1275 811 345
info@theIDLgroup.com
www.theIDLgroup.com

Contents

Glossary of terms used in this report

Attributes	Programme 'attributes' refer to any characteristic of a project or programme which is thought to be potentially meaningful and which would be useful to record systematically.
Macro-evaluation	An evaluation of a large set of projects by a range of methods. These may include meta-evaluations, evaluation syntheses, specially commissioned evaluations, desk analyses of existing data, etc, and ideally including an element of quality assurance of source data. Can be designed as a snapshot or as a sequential process.
Meta-analysis	Methods focused on contrasting and combining results from different studies.
Meta-evaluation	An evaluation of methodologies used by evaluations. Sometimes confused with synthesis of results of those studies.
Policy relevance	The extent to which interventions contribute to policy objectives as outlined in policy documents.
Project	Discrete interventions funded by DFID, as distinguished from a collection of interventions, such as country, regional or thematic *programmes*.
Programme	A collection of projects, such as a country programme, or regional or thematic programme.
Scheduled evaluations	Evaluations already planned and/or commissioned by DFID country offices or departments.
Specially commissioned evaluations	Evaluations that could be commissioned specially by the Policy Division as part of the process described in Next Steps.

Acronyms

AR	Annual Review
BAR	Bilateral Aid Review
BC	Business Case
CSO	Civil Society Organisation
DFID	Department for International Development
E&A	Empowerment and Accountability
EvD	Evidence and Evaluation Department
EQ	Evaluation Question
GNP	Gross National Product
GPAF	Global Poverty Action Fund (GPAF)
GTF	Governance and Transparency Fund
G&W	Girls and Women
LF	LogFrame
MDGs	Millennium Development Goals
ME	Macro-evaluation
PCR	Project Completion Report
PIMS	Policy Implementation Marker System
PPA	Programme Partnership Agreement
RCT	Randomised Control Trial
RED	Research and Evidence Division
SDA	Social Development Adviser
SIDA	Swedish International Development Cooperation Agency
SVG&W	Strategic Vision for Girls and Women
ToC	Theory of Change
TOR	Terms of Reference
USA	United States of America
VfM	Value for Money

Executive Summary

This report documents the findings of an evaluability assessment commissioned by DFID's Policy Division (with technical inputs from the Evidence and Evaluation Department). The evaluability assessment was carried out by *theIDLgroup Ltd.* between March and June 2012. It was undertaken in anticipation of proposed macro-evaluations of two DFID policy areas: empowerment and accountability (E&A) and the DFID Strategic Vision for Girls and Women (SVG&W). The purpose of the macro-evaluations is to assess and draw lessons from the wide variety of interventions being funded under each policy area.

The main *approach* to the macro-evaluations was originally envisaged by DFID as being one of *synthesising* findings from existing or scheduled evaluations, having first assessed the quality of data provided in these evaluations by way of a *meta-evaluation*. However, it was anticipated that some additional evaluations would need to be commissioned specifically as part of the macro-evaluation process, to look at evaluation questions unlikely to be answered by existing (or scheduled) evaluations. The evaluability assessment team was asked to assess the extent to which such additional work would need to be commissioned, as well whether or not a synthesis of wider research should be included within the macro-evaluations (see section 1.3)

The purpose of the evaluability assessment was to *"clarify evaluation questions; assess complexity and evaluability concerns; assess relevant evaluation approaches; consider budget implications; assess availability of data sources; and set out timeframes and milestones"* for the proposed macro-evaluations.

The methodology adopted by the evaluability assessment team was exploratory and iterative. It involved the development of potential tools for sampling, policy relevance assessment, and identification of testable hypotheses that could be used by subsequent macro-evaluations. Over 80 available documents for recently commissioned projects across seven[1] of DFID's 28 core countries were reviewed for potential evaluability and policy relevance. The stakeholders consulted as part of our work were the core DFID staff responsible for commissioning the evaluability assessment, and a number of colleagues with whom they work closely (see Annex B and Annex C).

The evaluability assessment found considerable gaps in the data that would be available for the macro-evaluations. While DFID has made an important and welcome commitment to public access to information on its expenditure and activities via its website, many key documents for current projects are not yet available on the website[2]. This is a major concern because it prevents any representative description and evaluation of DFID project activities. Evaluation reports are particularly difficult to find via website searches (see section 3.1).

The evaluability assessment proposed that attempts to synthesise achievements and lessons from interventions funded under the two policy areas should focus on DFID bilateral programme spending (which represents 37% of overall spend), rather than including spending through multilateral agencies, global funding mechanisms, or global and regional programmes. These have separate management structures and their own evaluation processes. Findings from the results of other such

[1] All projects funded since January 2011 in Ethiopia, India, Malawi, Nepal, Nigeria, Sudan, and Zambia,
[2] *projects.dfid.gov.uk*

evaluations should be used as a comparator, where there are similar policy objectives (see section 2.1).

Evaluation efforts should be concentrated on DFID's 28 "focus" countries, identified as a result of the 2011 Bilateral Aid Review (see section 2.3). Primary attention should be given to projects initiated from 2011 onward, as this is the year in which new and important policy commitments were made in respect to E&A and SVG&W. However, as many of these projects will be extensions or second phases of previous interventions, some impact data will be available relatively soon, and some comparisons will be possible between pre- and post- 2011 projects. This will enable DFID to respond to important questions about the timeframes required to effect change, particularly in the area of empowerment (see section 2.4).

Within the total population of post-2011 projects (and any previous phases of such projects) within the 28 focus countries, the macro-evaluations should focus on initiatives that are relevant to the objectives of each policy area. Identifying policy relevant interventions for each policy area is currently difficult, as the internal categorisations in use (input sector codes) do not readily capture all policy relevant projects (see section 2.4). A process for identifying the policy relevance of interventions would need to be developed and carried out as a preparatory step for any macro-evaluation. The process developed and applied by the evaluability assessment team (see section 2.4 and Annex B) could be adapted for use. The instrument for the scoring of projects would need to be adapted if it was also expected to identify mainstreaming of policy objectives (see section 2.4).

The evaluability assessment found that the main macro-evaluation approach initially envisaged by DFID, i.e. appraising and then synthesising the data available in scheduled evaluations, should not be the starting point for evaluations of the policy areas. The concerns are that the set of evaluations likely to be available is representative of a currently unknown population of projects; that they are too diverse and unrepresentative to enable the development of generalisations about findings; and that there will be a substantial delay before enough comparable quality evaluations are available (see sections 2.5 and 3.3.). The alternative is to use projects themselves as the primary unit of analysis, as the projects database provides an almost complete population of projects. Lessons from these projects could then be drawn through analysis both of mandatory project documentation *and* any existing or scheduled evaluations. Subject to the availability of project documentation, clusters of projects can be deliberately selected and worked with (see section 2.5).

The commissioning of syntheses of wider research is not recommended as a component of the macro-evaluation process. The conclusion of the evaluability assessment team is that a requirement to do so would substantially widen the scale of work involved, while at the same time making the boundaries of the enquiry less clear. In addition, DFID's Research and Evidence Division already has responsibility for commissioning evidence reviews. It would be more appropriate to make sure that their products were made use of as important supplementary sources, alongside the core data about policy relevant projects.

The evaluability assessment looked at possible ways of defining and mapping key attributes of policy relevant projects, which would be important for identifying clusters of interventions that could be the focus of evaluations. Attributes were looked for in the core project documentation that are most readily available online via the DFID projects database (Business Cases, LogFrames and Annual Reviews).

Within the SVG&W policy area it was possible to find clusters of policy relevant projects that share the same *outcome and/or impact indicators* and are thus comparable. This is likely to be more difficult with some E&A projects and requires the construction of additional measures to enable comparability. Categorising projects according to *types of interventions*, gleaned from key documents, was more challenging. Some interaction may be needed with project managers. Attributes describing project contexts were the most challenging to identify. Business Cases can help but require careful reading. Participatory (card sorting) exercises can help generate context information at different geographic scales. (See section 3.2).

Consideration was given to the options appropriate for the analysis of available data. It was noted that the comparison and synthesis of the experience of a collection of policy relevant projects will require a systematic and transparent procedure, if the results are to have any form of credibility and uptake (see section 3.3). Methods which can use nominal and ordinal scale data will be able to make the widest use of available evidence. A set of such method exist, which can produce comparable and testable results, which allow both theory and data driven approaches, and which can be participatory or expert driven.

The draft list of evaluation questions provided in the TOR for this contract was appraised for their likely evaluability. Interviews with key stakeholders provided further understanding of the relative priority of these questions, as well as the identification of potential gaps in the current list of questions. Questions were grouped into 7 different types, and consideration given as to the likely evaluability of each set of questions and the methods most appropriate to addressing them. Some of the groups of questions might be addressed through a synthesis of data available via mandatory project documents and availably evaluations. Others, however, would require specifically commissioned studies (see section 5).

As a result of the analysis summarised above, the evaluability assessment concluded that neither the E&A policy area nor the Strategic Vision for Girls and Women is yet ready for a macro-evaluation due to major gaps in available documentation (see section 3.1) and systemic difficulties with identifying investments in each policy area (see section 2.4). It is recommended that steps be taken to address these two issues as a matter of priority in order to address concerns for accountability.

Once data availability and policy relevance issues are resolved, it would be possible to design and commission evaluations to address many of the questions that DFID would like answered by the proposed macro-evaluations (see section 5).

Several of the questions could be answered by way of *synthesising* data already available to DFID. However, whereas DFID's intention had been to use existing or scheduled evaluations as the primary unit of analysis for such a synthesis exercise, the evaluability assessment identified several problems with this approach (see section 2.5 and 3.3).

Section seven of the report proposes an alternative approach. The proposed process addresses the joint purposes that DFID has for its proposed macro-evaluations, i.e. *accountability* and *learning*. The process would involve the customisation and refinement of evaluation questions for each cluster of projects for which synthesis of available data is proposed, and for any specifically commissioned evaluations. The process would also include periodic updating of the policy level ToCs to make them more robust and evidence based, incorporating the emerging evidence developed as part of the

process. While the proposed process addresses the *key objectives* of the macro-evaluations envisaged by DFID (as described in the evaluability assessment TOR), there are some key differences in the proposed *approach*. These are:

1. Priority is given to collation of basic descriptive information and the setting of boundaries through identification of policy relevant projects, before the use of any evaluations.
2. Policy relevant projects are the primary unit of analysis. Analysis of mandatory project documentation would be used alongside scheduled evaluations to seek answers to evaluation questions.
3. Rather than a snapshot approach of generating lessons, the process is iterative, allowing for a build up of knowledge over time.
4. Improved policy area ToCs are seen as a *product* of the evaluation process, rather than a *pre-requisite* to commissioning evaluations.
5. There is an annual reporting cycle rather than an interim and final report.

There are three kinds of expected *outputs* described in the next steps model:

1. The first is a description of the portfolio of policy relevant projects. The public availability of this information alone can be an important form of accountability, in addition to being an essential basis for subsequent planning and evaluation activities.
2. The second is a proposed annual synthesis report, using knowledge from recent evaluations, analysis of available project data, and external evidence sources. Reporting should be coordinated with the DFID corporate reporting cycle.
3. The third would be an update of the policy area ToC and its associated evidence base.

The proposed process has six key activities needed to reach these outputs; some of these activities will need to be done internally by DFID and some could be contracted out. Table 2 below summarises the steps proposed, the extent to which they can be contracted out, and the implied time-frames

Given the apparently high level of overlap in projects that are of relevance to both the E&A and SVG&W policy areas (see section 2.4), it is recommended that work commissioned to address the data issues and to prepare for subsequent evaluation work (steps 2-5) be managed jointly by the two policy teams. Some of the specific evaluations required (step 6) might be commissioned by one or other policy team, whereas others might be best managed jointly.

Step	Internally (by whom) or contracted out	Time-frame
1. Improve the coverage of project documentation.	Internal, although not the direct responsibility of either the Policy Division or EvD.	Coverage will be gradually improving over time. Other steps can start immediately, but the lower the coverage of documentation available on the website, the greater the work for DFID policy division and country office staff to locate documents.
2. Build a database of policy relevant projects started in 2011 or later.	Ideally in-house. If contracted out, the work should be done in close collaboration with DFID.	Can start immediately.
3. Identify clusters of policy relevant projects	Contracted out in close collaboration with DFID	Relies on step 2 having been completed.
4. Develop testable views of projects within clusters of comparable projects	Contracted out in close collaboration with DFID	Relies on steps 2 and 3 having been completed.
5. Use scheduled evaluations to test and analyse views	Contracted out in close collaboration with DFID	Relies on steps 2, 3 and 4 having been completed.
6. Use commissioned evaluations for special purposes **if** required:		
• Policy implementation review for SVG&W	Contracted out	Could start immediately (see section 5.1.1).
• SVG&W Pillar evaluations – snapshot view	Contracted out	Rely on steps 2 and 3 having taken place. Note that if there is no urgency in timing for pillar evaluations, the whole process described in this report should replace the need for separate for pillar evaluations (see step 6, section 7).
• Desk analysis of PCR ratings, risk and spend	Contracted out	Relies on step 2 having been completed (see section 5.1.4)
• Intervention to look at 'interaction effects'	Contracted out	Relies on step 2 having been completed (see section 5.1.2)

Table 2: Summary of next steps, management arrangements, and time-frames.

1. Introduction

1.1. Background

This report documents the findings of an evaluability assessment carried out by *theIDLgroup Ltd.* on behalf of the UK Department for International Development (DFID) between March and June 2012. The evaluability assessment was commissioned by DFID's Policy Division (with technical inputs from the Evidence and Evaluation Department [EvD]), in particular the teams responsible for two of DFID's key policy areas: the Strategic Vision for Girls and Women, and Empowerment and Accountability.[3]

These two policy areas were both the subject of renewed commitments by DFID in 2011.DFID has also undertaken to evaluate the impact and implementation of these two policy areas by 2015/16, at the end of the current Spending Round. DFID had considered doing this via separate or joined macro-evaluations[4] of DFID-funded initiatives of relevance to each of the policy areas. As a first step, DFID commissioned an evaluability assessment to address complexity and evaluability concerns, including the availability of data sources, budget implications, relevant evaluation approaches, and evaluation questions. It was envisaged that the evaluability assessment would lead into the design and then implementation of the two proposed macro-evaluations. The Terms of Reference (TOR) for the evaluability assessment are provided in Annex A.

The methodology adopted by the evaluability assessment team to address each scope of work covered in the TOR was exploratory and iterative. It involved the development of potential tools for sampling, policy relevance assessment, and identification of testable hypotheses that could be used by subsequent macro-evaluations. Available documents for recently commissioned projects across seven[5] of DFID's 28 core countries were accessed and reviewed for potential evaluability and policy relevance. The stakeholders consulted as part of our work were the core staff in DFID's Policy Division and Evidence and Evaluation Department responsible for commissioning the evaluability assessment, and a number of colleagues with whom they work closely, including the Vision for Girls and Women pillar leads. The methodology is described in detail in Annex B and a full list of stakeholders consulted is provided as Annex C.

1.2. What is an evaluability assessment?

An evaluability assessment has been described as a method for determining:

> *"whether the programme is ready to be managed for results, what changes are needed to do so, and whether the evaluation would contribute to improved programme performance"*[6]

Evaluability assessments involve judgements about: (a) the technical possibility of evaluating a programme and (b) about the practical value of doing so.

[3] Throughout the remainder of this document, these policy areas will be referred to respectively as E&A and SVG&W.
[4] See section 1.3. for a description of what was meant by a 'macro-evaluation'.
[5] All projects funded since January 2011 in Ethiopia, India, Malawi, Nepal, Nigeria, Sudan, and Zambia,
[6] Shadish et al. 1991 p.225

Evaluability assessments were first used in the USA by Wholey and colleagues in the late 1970s in response to the lack of use of evaluation studies. In the last decade a small number of evaluability assessments have been commissioned for a variety of planned evaluations of development aid programmes. Annex D provides a short list of relevant publications on evaluability assessments, concerning both methodology and practical applications.

The available literature is generally focused on the evaluability assessments of specific projects or programmes - not large undefined sets of programmes, as required of the evaluability assessment documented here. The literature also suggests that the boundaries between an evaluability assessment and evaluation design can easily overlap. Evaluability assessments should raise important questions that need to be answered by evaluators when developing an evaluation plan, but should not specify particular solutions to be used by the evaluators.

It is worth noting that a review of evaluability assessments in the 1970s and 1980sfound that they were rarely followed by an evaluation. However, it was also noted that *"the act of conducting an evaluability assessment in and of itself may provide enough guidance and programme attention in some instances to replace a more thorough evaluation*".[7]

1.3. What is a macro-evaluation?

By macro-evaluation, DFID refers to a process for assessing and drawing lessons from the wide variety of interventions being funded under each policy area. As outlined in the evaluability assessment TOR (Annex A, page 2), such a process would have two key objectives: *accountability* (what have been the effects from DFID-funded interventions?) and *learning and evidence* (what works and what doesn't, and how does context matter?).

The main *approach* to these macro-evaluations was envisaged as being one of *synthesising* findings from existing or scheduled evaluations, having first assessed the quality of data provided in these evaluations by way of a *meta-evaluation.*[8]

The rationale for suggesting such an approach seemed to be primarily financial: to commission additional evaluations specifically to look at the sorts of questions of interest would be costly, whereas using evaluations already budgeted for elsewhere was felt to be most cost-effective. It was acknowledged, however, that there *may* be a need to commission additional evaluations specifically to answer evaluation questions unlikely to be answered by existing (or scheduled) evaluations. The evaluability assessment team was asked to assess the extent to which such additional work would need to be commissioned, as well whether or not a synthesis of wider research should be included within the macro-evaluations.[9]

In this report the term macro-evaluation is used in the widest sense, to describe the nature and scale of the ambition, rather than a process using any specific set of methods or designs.

1.4. The Scope of Work and report structure

The scope of work documented in this report went beyond that required of typical evaluability assessments. According to the TOR for this assignment (see Annex A), DFID required the

[7] Leviton et al. (2010) Annual Review of Public Health 31:213-233. Downloaded from www.annualreviews.org
[8] See "DFID's approach to macro-evaluation", on page 2 of the TOR, provided in Annex A.
[9] See "Whether supporting work is needed", on page 4 of the TOR, provided in Annex A.

assessments to *"clarify evaluation questions; assess complexity and evaluability concerns; assess relevant evaluation approaches; consider budget implications; assess availability of data sources; and set out timeframes and milestones*".

The report structure broadly follows the requirements of the scope of work as described in the TOR. Sections two to four address programme attributes, data availability and Theories of Change. Section five examines evaluation questions. Sections six and seven look at future options and include discussion of timeframes, budgets and management issues. Under each sub-section heading, the particular scope of work addressed by that sub-section is cited in italics.

2. The units of analysis and boundaries of enquiry

Define the unit(s) of analysis for each evaluation and draw boundaries around the scope of programmes to be included…..assess to what extent the evaluations will ensure coverage across programme types and attributes.

Key points

- Attempts to synthesise achievements should focus on DFID bilateral programme spending (which represents 37% of overall spend), rather than including spending through multilateral agencies, global funding mechanisms, or global and regional programmes. These have separate management structures and their own processes for evaluating overall achievements Findings from the results of other such evaluations should be used as a comparator, where there are similar policy objectives. Evaluation efforts should be concentrated on DFID's 28"focus" countries, identified through the 2011 Bilateral Aid Review.
- Within the DFID focus countries, attention should be given to projects initiated from 2011 onward, as this is the year in which new and important policy commitments were made in respect to E&A and SVG&W.
- As many as a third of all projects initiated in 2011 or later are in fact extensions or second phases of previous investments. This means some longer term impact data will be available relatively soon, and some comparisons will be possible between pre- and post-2011 projects.
- Identifying policy relevant interventions for each policy area is currently difficult, as the internal categorisations in use (input sector codes) do not readily capture all policy relevant projects. This leads to significant data gaps.
- A Policy Relevance Rating scale, such as that developed by this evaluability assessment, could be used to assess the policy relevance of interventions. While an on-going "policy relevance ratings" exercise could be outsourced, there must be a process for validation by DFID staff.
- Evaluations should not be the primary unit of analysis for a macro-evaluation. Evaluation reports will however be important sources of information, alongside mandatory DFID project documentation.

2.1. Why explicit definitions of units and boundaries matter

Because of the global scale of DFID's development activities, some form of sampling is necessary before data can be analysed and conclusions drawn about those activities. Samples can easily suffer from selection bias, i.e. being unrepresentative of the population from which they are drawn. Selection bias can arise by accident as well as intention. Clarity about the nature of the population[10] from which a sample is drawn can make any sample biases more visible and manageable during analysis. Clarity about the boundaries of the data set should also enhance credibility of findings, because it is clear that the data can be re-examined by others, via a new sample if necessary.

[10]Used in the statistical sense. A population is any entire collection of people, animals, plants or things from which we may collect data. It is the entire group we are interested in, which we wish to describe or draw conclusions about.

2.2. Where DFID spends its money

DFID divides its overall budget into three broad categories: *multilateral assistance*, *bilateral assistance*, and *administration*. In 2010/11, 42% of DFID's budget was spent on multilateral assistance, 55% on bilateral assistance, and 3% on administration.[11]

The *multilateral assistance* budget covers the funding that DFID provides to international agencies such as the European Commission, the World Bank, the United Nations agencies, and the Global Fund for AIDS, TB and Malaria.[12] DFID's influence over, and ability to track, the way its money is used by these international organisations is limited. Furthermore it is not always easy to map the work of these organisations directly on to DFID's own policy priorities. Consequently it does not seem useful to include evaluations of multilaterals work directly into a macro-evaluation exercise that also includes DFID's own directly funded development activities. These organisations may, however, produce evaluations and research reports that are relevant sources of evidence for the development of E&A and SVG&W Theories of Change.

Of the 55% of the budget spent on *bilateral assistance*, 67% is spent by DFID country programmes. A further 8% is channelled through centrally managed mechanisms such as Programme Partnership Arrangements (PPAs), the Global Poverty Action Fund (GPAF) the Governance and Transparency Fund (GTF), or the Girls Education Challenge Fund.

Each fund has its own performance assessment framework and evaluation strategies applying to both individual grantees and to the funding mechanism as a whole. Some of these have generated synthesis reports which are relevant to the proposed macro-evaluations.[13] Others, such as the PPA and GPAF, will produce such synthesis reports over the course of the next two years. Because of the diversity of the projects funded via these mechanisms, the independence of their management and existing plans to synthesize evaluations of individual projects funded under them, it does not seem appropriate to include evaluations of their work alongside of those interventions directly managed by DFID and to then seek to synthesise the results as a whole. It would be better to use these independent syntheses as useful comparators, which may or may not support the findings of a macro-evaluation of DFID bilateral projects.

The main focus of this evaluability assessment report is on projects funded by DFID's bilateral aid programmes. This group of projects represent approximately 37%of DFID's annual budget.

2.3. Which countries?

The *projects.dfid.gov.uk* website provides documentation on DFID projects in at least 103 countries. In 2010 the bilateral aid review recommended reducing the number of countries DFID was currently working from 43 to 28"focus" countries. This is a relevant population for the proposed macro-evaluations, given that important new policy commitments were made this year for both E&A and SVG&W.

[11] http://www.dfid.gov.uk/About-us/How-we-measure-progress/Aid-Statistics/Statistics-on-International-Development-2011/Key-Statistics/

[12] http://www.dfid.gov.uk/What-we-do/How-UK-aid-is-spent/how-we-decide-where-aid-is-spent/

[13] Final Report-Learning from DFID's Governance and Transparency Fund(2010)
For future related events and publications see http://www.dfid.gov.uk/Work-with-us/Funding-opportunities/Not-for-profit-organisations/Governance-and-Transparency-Fund/GTF-learning/

If the primary aim of a macro-evaluation is to identify lessons to be learned, then the sample of countries should have maximum diversity. Because diversity exists between countries and within countries, a stratified sample would be appropriate. If, on the other hand, the aim is mainly to provide accountability, then the sampling of countries should be stratified by their share of the DFID budget.

In this evaluability assessment the sample of countries that was selected for examination sought to maximise diversity.[14] As described in Annex B (Methodology) the population of 28 countries was examined using a participatory Hierarchical Card Sorting exercise with DFID policy division and EvD staff. This process was designed to identify participants' view of the most significant differences between these countries, in terms of their consequences for project management and results. The process is described in Annex E and the results in Annex F. Eight sub-types of countries were identified via sorting exercises with each policy area team. A quota sample of seven countries was then selected, which represented 12 of the 16 sub-types. As well as guiding the search for relevant projects and documents, discussed in section 2.4 below, this process also generated potential evaluation questions, discussed in section 5. The same sampling process could be used as part of the design of a macro-evaluation and for sampling projects within individual countries.

DFID country programmes are a potential unit of analysis for a macro-evaluation. They would be one means of identifying *interactions* that are expected between projects within a country[15]. Such as those that might be described in a DFID country level strategy which spells out how the programme is expected to work as a whole. However country programme evaluations are no longer mandatory. The amount of this kind of evaluation material that will be available to review from 2011 onwards is likely to be smaller, unpredictable and take some time to become available.[16] Other ways of identifying interaction effects have been identified in section 5 below.

2.4. Which projects?

At the start of this evaluability assessment it was suggested that the focus of the planned macro-evaluations should be on investments started since January 2011, to be broadly in line with the introduction of the two policy areas.[17]Such a distinction was felt to be useful for two reasons: Firstly, to ensure that the macro-evaluation focused on investments that could be expected to be influenced by guidelines surrounding the two policy areas. The second reason was to put parameters around what would otherwise be a huge sample of potential projects to be looked at both by the evaluability assessment and any subsequent macro-evaluation.

However, some DFID staff felt that by excluding projects started prior to 2011, a macro-evaluation would risk losing the opportunity to learn from a number of projects which had begun prior to that date and which had been in operation long enough to have started producing results in areas relevant to the two policy areas. However, of the post-2011 projects that were looked at by the evaluability assessment team and found to be policy relevant, over one-third were of projects which were extensions or second phases of previous investments. If documents associated with those

[14] Diversity can be measured using three variables: Variety, Balance and Disparity (Stirling, 2007)
[15] A priority area of interest for both policy teams, see section 5.1.2.
[16] Up to 2010 DFID had a rolling programme of Country Programme Evaluations with 5 or 6evaluations of countries or regions per year, and each evaluation covering a five year period.
[17] The SVG&W was launched in March 2011. In Feb 2011the DCP endorsed increased emphasis in E&A.

projects are included in a macro-evaluation, information about outcomes and impacts, and the time-frames necessary to achieve change at this level, is likely to be available sooner than would be the case if only documents generated post-2011 are included.

As of May 2012, DFID had 612 operational projects in the 28 focus countries, of which 229 new projects were started in 2011, and more will come on-stream in each successive year. The challenge is to identify which of these are relevant to the policy objectives of the SVG&W and/or E&A policy area. There are no usable internal categorisations such as a Policy Implementation Marker System (PIMS) marker, and the OECD/DAC categories (input sector codes) used on the *projects.dfid.gov.uk* site are not helpful. Any attempt to synthesise knowledge of what has been achieved across a range of what appear to be policy relevant projects will be easily challenged by questions about selection bias, because the total population will be unknown.

Given that existing systems within DFID cannot easily identify which investments are policy relevant to each policy area, part of the work required of the evaluability assessment team was to develop an approach for doing so (described in the methodology section provided at Annex B).

Analysis of the policy relevance rating exercise carried out by this evaluability assessment found that 32% of projects in the sampled countries (where documents were available) were E&A policy relevant and 38% were SVG&W policy relevant[18]. This is equivalent to 47% of all projects being relevant to one or other policy area, and 24% of all projects being relevant to both policy areas. However, it is important to note that these percentages are suggestive only, because the number of projects is small (16 of 35 projects screened with available documents).

If the problem of identifying policy relevant projects can be addressed systematically, then projects are a feasible unit of analysis. Policy relevance rating could be outsourced but any results would need to be validated by DFID staff. The rating exercise could be adapted to differentiate projects where E&A or VG&W issues are the main focus versus being mainstreamed within projects with other policy objectives.

2.5. Which evaluations?

Unlike DFID projects, there is no mandatory listing of proposed evaluations in a DFID-wide database, so the total number planned for any period of time is not easy to establish. According to a database of scheduled evaluations set up by DFID Evaluation Department (EvD) in early 2012, there are approximately 340 evaluations scheduled over the next five years.[19]

These scheduled evaluations are many and varied. They include annual sector reviews, mid-term reviews, and ex-post evaluations; on-going evaluations run in parallel with monitoring systems and episodic evaluation; summative and formative evaluations; process and impact evaluations; thematic and country programme evaluations; joint evaluations and evaluations managed by DFID only; external evaluations and internal DFID peer reviews; evaluations led by host government, and by donors; focusing on DFID only projects, and others as well; using Randomised Control Trials (RCTs), analysis of survey data and participatory processes, action research, meta-evaluation of

[18] When 'policy relevance' ratings of 3 or 4 out of 4 were considered. See Annex B, Methodology, rating scale.

[19] The numbers vary across two Excel files sent to the evaluability assessment team by DFID, possibly because they were last edited on different dates. The files used are those dated 27 April (SVG&W) and 20 March (E&A) which were provided to the evaluability assessment team

contractors' self-evaluations, quasi-experimental, and ToC focused; and the issues they plan to address are also many and varied. A tabulation of some of these evaluation characteristics is provided in Annex G.

Of the 340 scheduled evaluations listed in the database, 116 (34%) of the evaluations were identified as policy relevant by DFID's EvD. Of these, 97 are of policy relevant projects in the 28 focus countries (60 E&A, 37SVG&W).[20]Of these the majority concern projects started before 2011. Nine will cover post-2010 E&A projects and 10 will cover post-2010 SVG&W projects, in 11 countries.[21] These represent around 14% of the policy relevant post-2010 projects as of June 2012.[22] It is likely that other evaluations of these new projects will be carried out, but information about them is not yet available.

As noted in section 1.3., DFID's proposed approach to the macro-evaluations was to use evaluations as the primary units of analysis, with additional special purpose evaluations being commissioned if required. More unusually, it was planned to begin this process before a set of relevant evaluations was completed. This could offer opportunities to influence forthcoming evaluations, in terms of the evaluation questions and potentially the methodologies used. If carried out iteratively, as each evaluation took place, there could be some cumulative learning and some knowledge would become available before 2016.[23]

There are, however, three problems with this approach. One is the diversity of the kinds of evaluations that have already been proposed, mentioned above, which will make it very difficult to develop any generalisations about the findings from such evaluations. The second is that evaluations are not mandatory and the sample of projects they do cover could easily be biased e.g. towards the more "successful" projects. The third is the incompleteness of knowledge available now in 2012 about the evaluations planned for post-2010 policy relevant projects. At this stage, information is available about evaluation plans for 14% of those projects. It seems likely that it will not be possible to know how representative (or not) the evaluations are of all DFID's work in the E&A and SVG&W policy areas for some years, by which time there may be limited opportunity to redress any biases in the kinds of evaluations used.

Because of these problems it would not be appropriate to see evaluations as the primary unit of analysis for the macro-evaluations DFID wishes to commission, even though they could be very important information sources. The alternative is to treat projects as the primary unit of analysis. Evaluations of any kind that examine that set of projects would then be used as one of the sources of evidence. Although evaluations are not mandatory, there are other kinds of DFID documents which are both mandatory and potentially useful: (a) Business Cases and LogFrames, (b) Annual Reviews (ARs) and (c) Project Completion Reports (PCRs). The latter are sometimes informed by, or are an actual sub-product, of mid-term and end of project evaluations respectively. In addition, both will contain achievement ratings of the Outputs and Outcomes (Outputs only for ARs), using the

[20]The other 20 being in countries outside the priority 28 countries, or else of regional programmes.
[21]26 others do not yet have a provisional timing.
[22] Being 47% (137) of a total 292 post-2010 projects, 292 is based on a search done on June 22, for projects started between Jan 2011 and May 2012. Actual figures found seem to vary week by week, perhaps as new data is added to the DFID database.
[23]The end of the reference period implied in page 1 of the TOR (Annex A).

improved rating system in operation since January 2011. All these documents should be considered as important evidence sources alongside any evaluations of the same project.

In addition to the scheduled evaluations taking place in specific countries there will also be evaluations of different funding mechanisms operating on an international or regional basis. Some of these will be of direct relevance, most notably the Governance and Transparency Fund, the Girls Education Challenge Fund, and selected PPA agreements. Because these funds have their own synthesis evaluation mechanisms, their most useful role will be as a comparator, against which results of evaluations of the two policy areas can and should be compared.

There have of course been evaluations of pre-2011 DFID projects in the two policy areas, and there are others scheduled for those projects that will complete in the next year or so. Reports on these should provide some information on strengths and weaknesses which could inform dialogue with designers of scheduled project evaluations (post-2010). (See step 5 of in Section 7 below)

2.6. Which research, what evidence?

The TOR for this evaluability assessment included a requirement to consider whether a synthesis of wider research should be included within a macro-evaluation process. The conclusion of the evaluability assessment team is that a requirement to do so would substantially widen the scale of work involved, while at the same time making the boundaries of the enquiry less clear. In addition, DFID's Research and Evidence Division already has responsibility for commissioning evidence reviews. It would be more appropriate to make sure that their products were made use of as important supplementary sources, alongside the core data about policy relevant projects. As will be explained in more detail below there will be a number of opportunities to do so, including when

- Assessing the policy implementation process within DFID (a particular concern of SVG&W policy area,), including the development of Business Cases[24]
- Assessing the validity of any assumptions made in the ToCs, both at the policy level and individual projects
- Considering the significance of findings reported in Annual Reviews and Project Completion reports and scheduled evaluations[25].

[24] Business Case requirements include "Strategic case A. Context and need for DFID intervention - Summarise relevant evidence underpinning the intervention"

[25] Formats require reference to evidence. See section "3. Evidence and Evaluation" for Annual Reviews and "Section C: Knowledge and Evidence" for Project Completion Reports.

3. Data availability and programme attributes

Examine existing data sources, including planned evaluations, and assess whether data generated is likely to meet macro-evaluation needs. Advise on whether additional data will be needed to enable comparison / generalisation and whether it will be built into evaluation design, e.g. thematic, cross-cutting and sectoral studies; or syntheses from wider research.

Key points

- The provision of public access to information and documents on DFID-funded development project via *projects.dfid.gov.uk* is an important and welcome development.
- Many key documents for current projects are not yet available on the website or via other searches by DFID staff. This is a major concern because it prevents any representative description and evaluation of DFID project activities.
- It was possible to find clusters of policy relevant projects that share the same outcome and/or impact indicators and are thus comparable. This will be more difficult with some E&A projects and requires the construction of additional measures to enable comparability.
- Categorising projects according to types of interventions, based on key documents, was more challenging. Some interaction may be needed with project managers.
- Attributes describing project contexts were the most challenging to identify. Business Cases can help but require careful reading. Participatory (card sorting) exercises can help generate context information at different geographic scales.
- Issues of data availability and identification of policy relevant projects will need to be addressed before undertaking any form of macro-evaluation covering numbers of policy relevant projects.

3.1. Documents

The *projects.dfid.gov.uk* website has been publishing documents relating to DFID projects since January 2011, as part of the coalition government's Aid Transparency Guarantee. As of June 2012 information was available on 1,767 completed projects, 1,512 operational projects and 102 planned projects, covering up to 123 countries. DFID estimates that 98% of all operational projects are listed on the website, with the remainder held back for reasons to do with "international relations, safety and security, personal information, commercially sensitive information". The most commonly available documents are Logical Frameworks, Business Case and Intervention Summaries and Annual Reviews. The website is being updated monthly, so documents relating to newly planned projects should progressively be included in the website.

The website is searchable using multiple attributes (both key words and options from set menus) and search results are downloadable in Excel. Some kinds of compound searches are possible; saving searches is not.

The database has the potential to be a very valuable resource for evaluators and researchers, contracted by DFID and otherwise. However, at this stage the database is far from complete. Key

documents are not available for the majority of projects, including those that have started in the last 18 months. The evaluability assessment team found the following number and percentage of documents currently available on the website:

	Up to and including 2010 (28 countries only)	**Post-2010 projects (28 countries only)**
Business Case and Intervention Summary	8% (27)	40% (108)
Logical Frameworks	22% (74)	39% (104)
Annual Reviews	17% (55)	9% (24)
Number of projects	100% (330)	100% (270)

Table 1:*Availability of project documents on the DFID website as of June 2012*

Locating documents not available publicly is not easily facilitated by DFID systems: only five out of 18 documents for projects under review and missing from the online database were located in other DFID systems. The size of the gap has implications for any attempt at evaluating a set of projects. It would be hard to establish how representative a sample is, if it is based solely on projects with available documentation. The set of projects with a full complement of key documents will be even smaller and thus more problematic.

According to the Evidence and Evaluation Department, senior management of DFID are aware of the weak coverage of the *projects.dfid.gov.uk* database and want to see it improved.

Evaluation reports are particularly difficult to find via database searches. They are not yet listed at *projects.dfid.gov.uk.* 276 are available via *http://www.dfid.gov.uk/What-we-do/Publications/Evaluation-studies/* but the search facilities are very limited, and only partly functional. The result of searches carried out by the evaluability assessment team were as follows: "empowerment" (7 reports), "accountability" (7 reports), "gender" (11 reports), "women" (7 reports), and "girls" (zero). *http://data.gov.uk/dataset/dfid-evaluation-reports*has no content at all. **Caveat**: The searches for project documents did not include direct enquiries to DFID country offices. This more labour intensive approach may have produced more documents, but it would not be a sustainable strategy for ensuring full coverage in the longer term.

3.2. Document contents

Programme attributes. Define the unit(s) of analysis for each evaluation and draw boundaries around the scope of programmes to be included. Define and map programme attributes; highlight implications for evaluation designs and assess to what extent the evaluations will ensure coverage across programme types and attributes.

Neither the E&A nor SVG&W policy team has a database which contains systematically collated data on the attributes of policy relevant projects.[26]This is a significant constraint on their ability to provide the most basic form of accountability – factual descriptions of the kinds of activities DFID is funding in these policy areas.

As mentioned in section 2.4 above, the projects listed on *projects.dfid.gov.uk* are not tagged according to the DFID policy objectives they address, or even internationally shared objectives such as MDGs. This is a surprising omission. The problem of identifying policy relevant projects is an immediate obstacle to the development of a usable database.

The evaluability assessment team experimented with ways to identify key programme attributes quickly through its analysis of programme documents (methodology in Annex B).Programme attributes were identified via:

1. Outcome and impact indicators (from LogFrames)
2. Descriptions & weightings of project outputs(from LogFrames)
3. Assumptions (from LogFrames)
4. Descriptions of theories of change and proposed interventions (from Business Cases)

Another possible attribute would be achievement ratings for outputs and outcomes, found in Annual Reviews and Project Completion Reviews. These were not looked at by the evaluability assessment team as only four Annual Reviews, and no Project Completion Reviews, were available for analysis (due to the post-2010 project threshold).

Achievement ratings provide a basis for comparing what could be quite different E&A or SVG&W projects in terms of their final results. Although they have their limitations, achievement ratings have these notable features: (a) achievement ratings have been mandatory for all projects, regardless of size, since January 2011, (b) although the rating system has been in operation since the 1990s, the recent revision of the rating system should provide better quality data, (c) outcome level ratings provided in Project Completion Reports are often a product of end-of-project evaluations. Similarly, but less often, output ratings may be a product of mid-term reviews.

Identifying attributes by outcome and impact indicators:

As described in section 2.4, 34 post-2010 projects were reviewed by the evaluability assessment team across the seven countries looked at. Of these, 15 projects were found to be 'policy relevant' to the E&A and/or SVG&W policy areas. Of these 15 projects, ten had LogFrames available to be reviewed. The evaluability assessment team collated and scanned the outcome and impact indicators of these projects to search for indicators that were used by more than one project. Nine such indicators were found. All ten projects shared at least one indicator with another project, and some shared two. This kind of data can be analysed relatively easily to find clusters of projects sharing the same one or more indicators.[27]

If a wider sample of DFID projects were looked at, it might be possible to find E&A and/or SVG&W projects that overlap in the specific outcome and/or impact measures they use. There may also be

[26] By attributes we mean any characteristic of a project which is thought to be potentially meaningful and which would be useful to systematically record.

clusters of E&A and or SVG&W projects that share similar outcome measures. If these clusters exist, it might be possible to compare project achievements in more absolute rather than relative terms (i.e. using achievement ratings).

This approach is more likely to find clusters of comparable projects in the SVG&W policy area than in the E&A policy area, because in areas such as health and education there are many outcome indicators that are widely used in similar forms. By contrast, in the E&A policy area the intangibility of objectives and difficulty in measuring the expected outcomes is widely recognised (McGee and Gaventa, 2011).In this policy area external evaluators looking at multiple projects may need to construct additional measures that could be applied across the projects they are examining. These could be in the form of a weighted checklist, participatory "success" rankings or multiple categories of outcomes.

Identifying attributes by the types of project intervention, from descriptions of project outputs

There will be some commonalities between sets of projects in the kinds of interventions involved, as described by the Outputs in the LogFrames. In theory projects could be tagged according to type of intervention, and clusters of similar projects then identified. Important as this could be, the evaluability assessment team's experience from looking at a sample of 34 LogFrames suggests this would not be a straightforward task. Ideally some prior work would need to be done to elicit appropriate categories of intervention types from relevant DFID staff. "Free card sorts" are perhaps the best way to do this, and can be done face to face or online.[28] Common types of intervention may explain similar outcomes, or signal the importance of other causal factors when the associated outcomes are very different.

Identifying attributes by project context: from Assumptions in LogFrames, or from Business Cases

Attributes describing project contexts will be the most challenging to identify and document. In theory they should be described in the Assumptions columns of LogFrames. In practice, in the cases examined by the evaluability assessment team, it was hard to identify specific aspects of local contexts that could make a difference to project outcomes. Quite the opposite, there were often default / pro forma type statements that could find a place in any project LogFrame. More useful information can be found in Business Case documents, but its extraction would be labour intensive. As with interventions, particular card sort exercises could be a useful means of eliciting perceptions of important differences in context. The results of the hierarchical card sortings of the 28 DFID focus countries, given in Annex F, show the kind of contextual differences seen as important by the two policy area teams. Similar exercises could be done with projects found within specific countries[29].

3.3. Analysis methods that suit the available data

Programme attributes. Define the unit(s) of analysis for each evaluation and draw boundaries around the scope of programmes to be included. Define and map programme attributes; highlight implications for evaluation designs and assess to what extent the evaluations will ensure coverage across programme types and attributes

[28] See How to Sort, by Harloff and Coxon, 2009 and their associated The Method of Sorting website.
[29] The caveat here is that the participants must have at least basic knowledge of the items in any set being sorted.

The comparison and synthesis of the experience of a collection of policy relevant projects requires some form of systematic and transparent procedure, if the results are to have any form of credibility and uptake. In the words of Michael Scriven (1994) *"The lack of explicit justification of the aggregation procedure is the Achilles heel of assessment efforts".*[30]Systematic reviews, including those recently funded by DFID, AusAID and others, are an attempt to address this challenge. These have involved a rigorous assessment of evidence, and demanded high statistical standards.

The downside of this approach to the syntheses of findings is that systematic reviews typically involve the rejection of a large proportion of studies not seen as amenable to statistical meta-analysis.[31] In the case of DFID project-focused evaluations this proportion could be very high.

This problem is especially relevant to attempts to synthesise the results of evaluations of projects concerned with empowerment and accountability, where measurability of outcomes is an acknowledged challenge, and scientific experimental approaches are few and far between. Where the SVG&W is also concerned with empowerment it will also be faced with this issue. This is more so in the 'foundation' and the 'roof' of the house, and less so in the pillars.

One possible response is to revisit ideas about appropriate levels of measurement (i.e. nominal, ordinal, interval, ratio scales) that can be used in evaluations and evaluation syntheses. More demanding levels of measurement are not ideal, if they eliminate the use of large areas of project experience. The alternative is to find forms of analysis that can use nominal scale data (i.e. categories). This kind of data is easily obtained and will enable evaluators to make use of the widest range of data sources. Higher measurement levels (e.g. ordinal, interval, and ratio level measures) can be simplified down to categories,[32] but not the other way. Aspects of the context, interventions and outcomes can all be described using binary categories, and whole projects can be described using multiple sets of categories. As with the use of more sophisticated measures, care must be taken to carefully apply such categories[33] .

Annex H provides more detail as to how this can be done using a range of methods, which produce testable results.

3.4. Additional data via evaluations and studies

Advise on whether additional data will be needed to enable comparison / generalisation and whether it will be built into evaluation design, e.g. thematic, cross-cutting and sectoral studies; or syntheses from wider research

The more immediate challenge facing any form of evaluation of the E&A or G&W policy areas is how to access the most basic forms of data, and carry out a basic analysis of that data, including:

[30]Scriven M. The Final Synthesis. Sage, American Journal of Evaluation, 15/3(1994):367-82.
[31]Davies, R, (2011) 3ie and the Funding of Impact Evaluation. A Discussion Paper for AUSAID.
[32]By using cut-off points to define two ends of a scale, which then become two different categories.
[33]To ensure construct validity (you are describing what you think you are describing) and measurement reliability(others use the category in the same way you do)

- Identification of which operational projects are policy relevant;
- Obtaining the mandatory documents required from these projects;
- Categorising and/or measuring the attributes of these projects (covering context, interventions and outcomes).

Until these tasks are addressed, commissioning any additional "thematic, cross-cutting and sectoral studies; or syntheses from wider research" should be a lower priority.

If and when clusters of project are identified which have shared outcome measures it is likely that there will be a number of scheduled evaluations that will examine at least some of the projects in that cluster. The timing of those evaluations and their coverage (i.e. percentage of all the projects in the cluster) will need investigation. Low coverage or long delays before an evaluation is scheduled may justify steps being taken to encourage project managers to carry out an evaluation, to do an evaluation in more depth or to do one earlier than planned.

4. Assessment of the Theory of Change

Assess existing theories of change, and provide recommendations for linkages and improvements, and in consultation with DFID, revised change models for the macro-evaluation. Examine opportunities for an 'evolving' ToC for the macro-evaluation, and suggest at what points and how ToC and Evaluation Questions could / should be revised.

Key points

- ToCs are more often developed for specific projects or programmes than for overarching policy areas, where a single coherent theory of what works is less easy to identify or depict.
- However, policy level ToCs can be useful to communicate the policy area, to set a direction for the future, and to provide a summary description about existing activities.
- A ToC should ideally evolve to reflect an accumulating body of evidence and testable knowledge about what works in what circumstances to deliver results in the project, programme or policy area it seeks to explain or depict. DFID's requirement that the macro-evaluation process should be used to update their policy level ToCs is therefore appropriate.
- The ToC in both E&A and SVG&W policy areas are currently unevaluable. This is recognised and steps are being taken to make them more robust and evidence-based.
- The most common problem needing attention is the lack of clarity about expected linkages between events described in the ToC. ToCs could then be made more evaluable, and informative, by identifying projects that exemplify linked events within each ToC.
- A policy implementation review of the Vision for Girls and Women (referred to in section 7) could test ownership of the SVG&W ToC within DFID, and extent to which it adequately communicates the policy direction of the Vision. Evolution of the policy area ToC should not be problematic. Connections between existing events can be "rewired" and new exemplar projects can be found for new events placed in the ToC. Unlike project level ToC, there will not be extra data collection costs, or loss of value from past data collection efforts.

4.1. What is a Theory of Change?

Funnel and Rogers[34] define a programme theory as *"an explicit theory or model of how an intervention contributes to a set of specific outcomes through a series of intermediate results. The theory needs to include an explanation of how the programme's activities contribute to the results, not just a list of activities followed by the results".* In many people's eyes this is also a workable definition of a Theory of Change (ToC).The evaluability assessment team's own even simpler version is *"a description of a sequence of events expected to lead to a desired outcome, which is verifiable".*

Theories of Change are most often developed to model the sequence of events and desired outcomes of a discrete intervention. They are not generally created to describe a collection of

[34] Funnel, S., Rogers, P. (2011)

diverse interventions all contributing to a high level policy objective where a coherent and testable theory about what works is less easy to identify or depict. However, such a high level ToC can be useful for the purposes of:

> **Communication**: to simplify a complex situation to help explain it to others and persuade them of the logic of the proposed interventions, or
>
> **Management**: to model a situation to better understand it and programme around it. Management functions could include: (a) to set a direction for the future, (b) to make a summary description about existing activities.

For both the E&A and SVG&W policy areas, DFID's Policy Division has drafted what DFID staff members variously refer to as 'change models', 'theories of change', and 'policy diagrams'[35] to communicate and model the inputs, sequence of events and desired outcomes of each policy area. The SVG&W has both an overarching ToC, known as 'the House', as well as stand-alone ToCs (in varying stages of development) for of each 'pillar' of the house.

Given the complexity and uncertainty of development interventions, a ToC should ideally evolve to reflect an accumulating body of evidence and testable knowledge about what works in what circumstances to deliver results in the project, programme or policy area it seeks to explain or depict. Rondinelli (1993) argued that *"continuous testing and verification is required if development activity is to cope effectively with the uncertainty and complexity of the development process"*[36].

DFID recognises the importance of an evolving ToC and requires the proposed macro-evaluations to include periodic revision of these ToCs[37]. Such an evolving ToC could be used both internally as a resource for those implementing the policy or programme, but also externally as both a public knowledge good and as a form of accountability.

Section 7 of this report, next steps, includes comment on how evidence emerging as part of the iterative evaluation process being recommended be used to update the ToCs on an annual basis.

4.2. Assessment of the current ToCs

Relatively little has been written on criteria relevant to the *assessment* or *evaluation* of a ToC. Funnell and Rogers (2011) have provided two lists of checklist type questions (pages 294-6). A set of criteria recently proposed by Davies[38] were used to assess the E&A and SVG&W Theories of Change.

4.2.1. Assessment of the E&A ToC

The E&A ToC is captured in a single page diagram, which is a hybrid of a stage and network model. The main purpose of drafting this ToC was to guide and link up a diverse field of interventions into a coherent strategic narrative and set of impacts.

[35] During interviews and workshops with the evaluability assessment team it was clear that there is not yet a single agreed label or purpose for these change models.
[36] Development Projects as Policy Experiments, Dennis A. Rondinelli, Routledge, 1993.
[37] See scope of work 4, TOR, Annex A.
[38] See http://mandenews.blogspot.co.uk/2012/04/criteria-for-assessing-evaluablity-of.html

The complexity of the model does not seriously weaken the extent to which it can be *understood*[39] by its key stakeholders. The events in the model are not easily *verifiable* in the way that events in a LogFrame are, via associated indicators, but it would be possible to develop indicators for many of the events especially on the right side of the model. There are identifiable and potentially *testable linkages* between the events in the model. However, there are not yet accompanying *explanations* of how the linkages work. DFID's E&A team have identified some pathways that need to include more intermediate events, to clarify the processes involved. The model is detailed at the beneficiary end but less *complete* at the other end where interventions begin. The E&A team have identified some event boxes at the beginning which they think need disaggregation into different events. The *inclusivity* of the model is yet to be tested. In its favour are the multiple causal pathways, allowing for different approaches which are inevitably necessary given the wide range of contexts in which E&A programmes have been established. *Plausibility* is difficult to establish from the diagram by itself. Identifying exemplar projects for the various parts of the causal chain would be helpful. *Ownership* of the ToC seems to exist within the E&A policy team (wider ownership was not tested as part of this evaluability assessment).

In its current form the E&A ToC is not evaluable. However, it is arguable whether a ToC that encompasses a third of DFID projects should be tested as a single coherent theory of what works, notwithstanding the realistic presence of multiple causal pathways. However, it should be valuable as a usable *map* of what DFID is doing, one which provides information about different ways of getting to where you want to go. But to be usable a map needs to be accurate, it needs to be clearly linked to the territory it represents. Looking at any part of the ToC (the map) can we find exemplar projects that show those processes working in real life? Looking up from the territory, can we find where any chosen E&A project fits within the ToC? Asking these questions is one relatively simple way of testing the usefulness of such policy level ToC. Individual project evaluations could go into more detail, by examining whether they function in the way that their place in the ToC suggests.

4.3. Assessment of current G&W ToC

The SVG&W ToC is a more complex modular hybrid model. The "house" contains four pillars, each of which has been developed into a separate subsidiary ToC, and these use different combination of tables, stages and network structures. The "house" model was developed with a concern for *communicability* and seems to work in that respect. There are verifiable measures of change at the pillar level and in one version of the Vision level.[40] Causal linkages between events are less clear than in the E&A ToC, but the SVG&W team have taken some steps to find projects that address multiple linkages and to document the linkages involved. This is a concrete example of the exemplar process suggested for the E&A ToC above.

The "house" model is not inclusive; it intends to focus on those elements which it sees as essential. This poses a risk in terms evaluability, as alternative approaches will not be so identifiable and comparable. There are efforts underway to collate evidence which justifies the approach being taken, but there is also considerable uncertainty about how important it is that all four pillars need to be addressed if the overall vision is to be achieved. That is seen as one question a macro-

[39] *Understandability* is one of Davies' ten criteria for the evaluability of a ToC, which overlap with many of those proposed by Funnell and Rogers 2011. Other words in italics in this paragraph represent other ToC evaluability criteria.
[40] Which refers to "reduced intergenerational poverty rates of girls and women" and "reduced discrimination against girls and women"

evaluation could address. Ownership of the house model is also seen as uncertain. This could be an issue explored by a specifically commissioned evaluation (or "policy implementation review") looking specifically at questions regarding implementation of the Strategic Vision (the option is discussed in section 5.1.1).

The Pillar ToCs are quite varied. Pillar 1 (Delayed pregnancy and safe childbirth) has the same format as the E&A ToC and essentially the same kinds of strengths and weaknesses, but with a greater need for disaggregation of events, and linkages between them, as shown at the input end of the model. Pillar 2 (Economic assets) is more rudimentary showing lists of events at five different stages of a causal chain, with no information on expected causal linkages between them. As such it is unevaluable. Pillar 3 (Girls education) seems to exist in text form only at this stage. It has the same limitations as Pillar 2, only more so, because the level of detail at each stage is much less. Pillar 4 is much more detailed, but with linkages that are essentially generic pointers rather than claims about specific causal pathways.

The comments made above about *expectations* of evaluability with regard to the E&A ToC also apply to the SVG&W ToCs. An examination of the SVG&W ToC also raises other important issues. It is important to clarify what elements within each of the ToC are *necessary* or even *sufficient*, to ensure the achievement of overall objectives. Knowing these answers has relevance for both programme design and the planning of evaluations to test a project or programme's ToC. The existence of multiple pathways in the E&A and Pillar 1 ToC suggests that there is no one necessary element. At the same time, because each of these pathways involve multiple different events, it is clear there is no single "silver bullet" that is a sufficient cause on its own. These are important claims, in a world where experimental approaches seem to be focused on finding silver bullet interventions. In other SVG&W ToC the causal links are not yet clear enough to identify where each theory stands. More clarity would help.[41] At the "house" level there is interest and uncertainty about the extent to which interventions involving a combination of two or more pillars are necessary, or whether any pillar alone is sufficient. There is also interest in exploring which dimensions of the 'enabling environment' are most powerful.

4.4. Evolving Theories of Change

A diagrammatic version of a ToC can be revised at any time at minimal cost (in terms of data collection costs) by "re-wiring" *the relationships* between existing events (and any associated indicators). This can be done by presenting possible relationships in an Excel matrix in a workshop setting with appropriate stakeholders.[42]Or, where they are distant, by constructing a multiple choice questionnaire that provides the same choices. This approach treats events in the ToC as the equivalent of building blocks whose relationships with each other can be dismantled and reassembled as needed. Ideally, the rationale for the changes being made should be recorded so there is an "auditable trail of intentions", from which lessons could be drawn by any evaluations at a later date.

Changing *the listed events* in the ToC at a project level is a more costly exercise, either because (a) they involve the introduction of new data collection costs and/or (b) they may involve the waste of

[41] For a graphic illustration of the differences between sufficient and or necessary causal conditions, see the different possible combinations at http://www.mandenews.blogspot.co.uk/2012/06/representing-different-combinations-of.html
[42] See this example from a DFID Indonesia maternal health project

past investments in data collection, if some old events are no longer seen as useful. Normally these kinds of change should be introduced less frequently and after more deliberation. However a ToC at a policy level may not be subject to the same kinds of constraints. As suggested above, events in a policy ToC can be anchored to reality by reference to specific projects where they can be found, rather than specific measures of those events. Redefining or adding new events in a ToC would then require finding new exemplar projects, which will already have their own indicators and data collection mechanisms.

There are other less avoidable costs arising from the time it takes to seek a consensus on any revised model. Given the relatively small size of the two policy teams this may not be a major concern. However, communicating versions to a wider range of DFID stakeholders would take time.

5. Evaluation questions

Define and recommend core evaluation questions for each evaluation. A suggested list of possible questions for each macro-evaluation is at Annex 1

Key points

- Questions relating to the relevance and implementation of the SVG&W are a high priority for DFID and could be addressed via a specially commissioned review using a combination of staff surveys and a desk study of project documentation.
- Questions relating to interaction effects of different interventions are of concern to both policy areas, and findings would have consequences for expected costs and effectiveness of future project designs. If data availability problems can be resolved, scheduled evaluations can be used to test some hypotheses built up from project documentation about a set of comparable projects.
- The value of mainstreaming versus specialisation of projects is of interest to both policy areas. Because of the more complex comparison problems involved, a "case study" via a scheduled country programme evaluation may be the most appropriate approach.
- Both policy teams have questions about overall impact. Some conclusions may be reachable using the new DFID project rating system, but they will require some progress with identifying policy relevant projects and types of these. Results will be useful mainly for accountability purposes. An analysis of the PCR ratings of E&A and SVG&W projects could be specially commissioned for overall accountability purposes.
- Opportunities for assessment of overall achievements using more absolute measures will be limited to two of the SVG&W "we wills" measures and one "house model" indicator (if collected).
- VfM questions are a priority for the SVG&W team. A crude VfM analysis will be possible using Project Completion Report data, relating costs to achievement levels. More in-depth analysis of costs per unit output and outcome will need to come from scheduled evaluations.
- Issues of how context matters and what works in what circumstances were a strong concern for the E&A team. Generalisations about the importance of particular contexts and interventions will be possible, but will require the development of good data sets about projects with comparable outcomes.
- 3ie's experience with the use of Systematic Reviews has highlighted the value of specific rather than broad evaluation questions. These will be easier to identify when the focus is on projects with common outcomes, and when there is a dedicated hypotheses building stage prior to an evaluation.
- Nevertheless, private sector experience indicates that where good data sets are available, more open ended enquiries can be used to find important associations between contexts, interventions and outcomes.

The extent to which the evaluability assessment could be expected to *'define and recommend'* core evaluation questions was a focus of discussion between the evaluability assessment team and DFID

at the start of this contract. The inception note prepared by *theIDLgroup* following the initial briefing meetings stated that the evaluability assessment would:

> "*help prioritise and refine the current list of proposed evaluation questions, to test the evaluability of these priority questions, to identify gaps, and on that basis to advise on whether additional studies, or components of studies, will need to be commissioned as part of the macro- evaluations*". It was also noted that "*Further validation and refinement of the evaluation questions will need to take place during the macro-evaluation design phase*."

The TOR for the evaluability assessment listed the sorts of evaluation questions that DFID is interested in. Annex I provides comments on improvements that could be made to each of these questions. Evaluation questions were further explored during interviews and workshops (the Card Sorting Exercise) [43]with key stakeholders as part of the evaluability assessment. In the section below, the evaluation questions noted in the TOR and/or elicited through interviews are grouped into six categories. These groups are prioritised, their evaluability assessed, and comment made on the type of study that would be required to answer each group of questions. Comment is also made on the gaps identified in proposed evaluation questions. The section concludes with arguments for two approaches to asking evaluation questions and anticipation of the kinds of explanations they can generate and how they could be presented.

5.1. Categories of evaluation questions

5.1.1. Policy implementation

Almost half of the 20 sets of questions listed by the SVG&W team in the evaluability assessment TOR refer to issues of policy relevance and implementation[44]. By this we mean the process connecting the original c*onception* of the Strategic Vision to the approval of projects which are seen to embody the new policy direction. Interviews confirmed that this is a priority. In summary, these questions asked about:

- Changes in the allocation of financial resources to programmes on girls and women
- Appropriate organisational structures for effective delivery of the Vision
- How the Vision guided the work of DFID Country Offices on girls and women
- Impact of the Vision on DFID's relationship with external partners

These questions are answerable, but a synthesis of other evaluations would not be the best means of doing so. The first category of questions could be answered via an analysis of documents available within DFID and the last three via surveys of DFID staff and external stakeholders by external consultants with policy development and evaluation expertise. This could be done by a specially commissioned evaluation (or 'policy implementation review') and without delay. Progress would not be dependent on resolution of the data availability issues raised earlier in this report.

5.1.2. Interaction effects

Both policy teams suggested questions about the importance of multiple combined interventions versus single interventions. The E&A team is interested to know whether interventions that combine

[43] See Annex B – Methodology.
[44] Question sets with arrowheads, which the evaluability assessment team has numbered 2,3,4,6,13,14,17,18,19.

both empowerment and accountability elements achieve better development results than those working solely on empowerment or solely on accountability[45]. The SVG&W team is interested in the effects of interactions between different pillars and whether interventions addressing more than one pillar achieve better outcomes for girls and women than single pillar interventions[46]. These are important and practically useful questions because single interventions may be less expensive and easier to deliver than multiple coordinated interventions. On the other hand, there is a belief in both teams that combined interventions are necessary to be really effective.

Evaluation of these questions is dependent on three kinds of conditions. Firstly, comparable sets of projects need to be found with each policy area, i.e. projects with comparable outcomes, where differences in interventions could potentially be visible. Secondly, within these sets there need to be differences in the extent to which projects try single versus combined interventions, but which are otherwise as similar as is realistically possible. Thirdly, there needs to be information about other differences between the projects, which could provide alternative explanations for differences in outcomes.

These conditions are often described as "natural experiments"[47].It is unlikely that these conditions will be found simply by searching for and synthesising results of existing and/or scheduled evaluations. A more proactive approach would involve the use of scheduled evaluations to test some hypotheses built up from existing project documentation about a set of comparable projects. These could: (a) verify that there were performance differences between the projects, and (b) check for alternative explanations, such as additional interventions being provided by third parties (e.g. other aid organisations or host governments). This approach would be dependent on sufficient data being available from project documentation to make informed predictions about differential outcomes, a problematic issue raised in section 3 above. If that cannot be resolved, then a commissioned evaluation would be needed, which would include initial data gathering about project and context differences. Even this approach would need some basic project data, from which to select a sample of relevant projects.

5.1.3. Mainstreaming

Both policy teams have posed questions about the value of "stand-alone" projects versus integration of their interventions into other projects.[48] Identifying these kinds of projects should be possible, with a revised policy relevance rating scale[49] of the kind used in this evaluability assessment. Identifying differences in outcomes will be more difficult than when examining interaction effects. The relative advantages of these two approaches are presumably depth of impact versus breadth of impact respectively. Depth of impact being expected with stand-alone projects and breadth via a wider set of "mainstreaming" projects. Both measures would need to be obtainable from both sets of projects. There may also be other parallel theories about these approaches which are relevant. For example, that it is easier to fund or implement one of these two approaches, and that this magnifies their benefits or compensates for their weaknesses (re breadth or depth). These are all important claims with programming implications.

[45] Question set 3 of 5 on page 7 of the TOR.
[46] Question set 10& 15 of 20, plus interviews with the gender team of the policy division.
[47] See http://en.wikipedia.org/wiki/Natural_experiment
[48] E&A question set 4, and SVG&W question set 18
[49] See Annex B - Methodology

Given the wider diversity of projects that would need to be compared and the causal complexity that could be involved, a natural experiment of the kind described above is unlikely to be workable. A country level case study would be more useful, at least as a first step. Scheduled country programme evaluations could be an opportunity for such a case study.

5.1.4. Overall impact

In the draft list of evaluation questions provided in the evaluability assessment TOR, both policy teams asked questions about overall impact[50].

The newly designed system for scoring project achievements in Project Completion Reports[51] should provide one means of making global statements about the overall achievement of E&A and SVG&W projects. The focus is now on actual rather than expected achievements and there is now a wider rating scale recognising more differences in performance. There have been analysis of project completion ratings by DFID in 2005 and 2010 and it is likely that there will be analyses of ratings under the new system, possibly by 2015. That analysis could provide potentially useful comparators (e.g. the average of all projects, or all projects from comparable countries or of similar scales).

Policy teams could engage with those responsible to seek a disaggregation by project type, using OECD input sector codes (at worst), or by tagging projects using their own policy relevance ratings scales (preferable). In the absence of plans for an analysis of PCR ratings, a simpler exercise could be contracted out, to focus on E&A and SVG&W projects only. In either case, an analysis of the PCR ratings of E&A and SVG&W projects would be useful for overall accountability purposes.

The E&A team have also asked more specific questions about overall impact, referring to particular kinds of outcomes (poverty, development, fragility, governance outcomes, or to social cohesion or power relations). Finding answers at this level would require additional tagging of projects by outcome type, preferably prior to any analyses of PCR ratings.

The new PCR formats will also provide some opportunities for global analysis of lessons learned. Relevant sections include section 1.5 and 5 (What lessons have we learned about what went well, including lessons that will affect future project design?) and section 3.1 (3.1 PCR – Assess any changes in evidence and what this meant for the project). Lessons learned were examined by previous analyses of PCRs and may be in the future. Doing so will more costly than analyses of scoring data, so the challenge will be to extract information that goes beyond truisms, and has some use. If analysis of project scoring was contracted out then this work could include scanning of contents for exceptional lessons that have or could make a difference.

The SVG&W team asked three questions about results on a global scale: *(a) to what extent are the results achieved by the vision quantifiable and measurable? (b) Did the vision achieve these results it set out to? (c) What impacts has the Vision had on empowerment of girls and women?*

Efforts have already been made to make them measurable, by identifying nine "we wills", mainly tracked through the Corporate Performance Framework[52].Five of these seem to be about DFID

[50] E&A question set 1 and SVG&W question set 7, 8,
[51] DFID How To Note - Reviewing and Scoring Projects, November 2011
[52] Minute to Michael Anderson, Sept 2011: Monitoring and Reporting Results for Girls and Women

supported activities, rather than changes in people's lives[53]. Two of the other three are readily measurable[54]. So the answer about measurability (at a global level) at this stage seems to be "to a limited degree only"

Attribution of changes in the measures to the Vision could be sought by looking for correlations between successful policy implementation (as discussed above) and results on the two usable results indicators at a project level. Where they are found attention would then need to be given to the significance of other possible influences, including other agencies and other policies. These could be explored by making use of scheduled evaluations of the projects examined. The same scheduled evaluations could look out for other unexpected / unmeasured impact differences between projects seen as more versus less successfully implementing the Vision.

The roof of the SVG&W "house model" includes two other global measures: (a) Reduced intergenerational poverty rates of women and girls, and (b) Reduced discrimination against women and girls. The first of these would not be evident within 4 years available for a macro-evaluation or its alternatives, but could be identified via a modest scaled tracer study or analysis of national survey statistics. Reduced discrimination could be operationalised via various measures, and tracked within a range of SVG&W projects. Attribution could be assessed by the same means as with the usable "we wills" measures.

5.1.5. Value for Money

Three sets of SVG&W evaluation questions relate to Value for Money. The first two ask whether VfM was achieved, and the third about differences in VfM between various approaches. Absolute answers are unachievable since there is no gold standard. Comparisons could be made with past policy periods, such as GEAP, if there was comparable VfM data available from that period. This does not seem to be the case.[55]

Comparisons between post-2010 policy relevant projects in terms of VfM are more feasible. A crude but potentially useful analysis could be made of the relationship between project cost and PCR ratings (a kind of cost-effectiveness analysis)[56]. Of special interest in terms of their potential learning value would be the outliers (high cost, low PCR achievement ratings & low cost – high PCR ratings).

More in-depth analyses could be made within clusters of projects with comparable kinds of outcomes, discussed in section 3 above. Scheduled evaluations could be used to collect detailed cost information. Comparisons could be made of output unit costs (efficiency). Although more challenging, it might be possible to put costs on per capita outcomes.

5.1.6. What works in what context

This kind of question has been asked by both policy teams, but more so by the E&A team.[57]The SVG&W team's interest in context has been largely focused on the influence of "the enabling environment". The E&A team has had a wider interest, captured in two sets of related questions

[53] E.g. "Number of children supported by DFID in primary education per annum" (underlining added)
[54] Relating to family planning methods and use of skilled birth attendants
[55] Gender Equality Action Plan (GEAP) Light Touch Review: Summary Document , DFID, 10 September 2010
[56] See distinctions between VfM terms here http://www.mandenews.blogspot.co.uk/2012/05/perspective-on-value-for-money.html
[57] E&A question sets 2 and 5; SVG&W question set 11

about context: how programmes work differently in different contexts and with different groups, and whether it is possible to generalise about kinds of contexts that make a difference. The importance of context in E&A initiatives is strongly argued by McGee and Gaventa (2011)[58], in their review of the impact of transparency and accountability initiatives ("Context as Crucial" p19).

The classic approach to finding answers about how context matters is to do in-depth case studies, either via research or customised evaluations. Their results should be immediately useful to local project management. The bigger challenge is how to address the E&A team's second interest – how to generalise about the importance of specific kinds of contexts.

Doing so will require some systematic data collection about the presence of different kinds of context features across projects with comparable outcomes. Some of these can be derived from the policy area ToC, e.g. by disaggregating the concept of "enabling environment". Other more inductive approaches use case comparisons, such as the card sorting exercise used in this evaluability assessment (focusing on projects rather than countries). Context differences should also be available from readings of project documentation, especially Business Cases. Information would also be needed on intervention differences, because they are expected to make a difference to outcomes. That information should be available, in the first instance, from project documentation.

It is possible to then do a desk based analysis of a data set with context, intervention and outcome information on a set of projects using methods explained in Annex H (especially QCA and Decision Tree software). However the results, in the form of association rules (i.e. IF this context AND intervention, THEN this outcome), need verification. Scheduled evaluations should provide opportunities to do so. They can be used to check if attributes assigned to each project were appropriate and whether causal mechanism thought to explain the associations can actually be found on the ground. If not, the associations could be spurious[59].

In a cluster of projects with common outcome measures, the various evaluations scheduled for the different projects in the cluster will provide a sequence of opportunities to test, refine then re-test analyses of "what works in what circumstances". This could generate more rigorous findings than a once-off synthesis of a set of evaluations.

5.1.7. Gaps? Risk and failure

In the list of draft evaluation questions provided in the evaluability assessment TOR there was only one explicit question about risk[60] and none about failure. Yet there are now more high risk projects than in the past[61], and in the past at least 20% of DFID projects failed[62]. Two general questions could be asked, using ARIES data in the first instance. The first is descriptive, about the incidence of high risk and failed projects, and would be useful for accountability purposes. The second is evaluative, about the relationship between risk and failure: (a) How correlated are these attributes? (b) What causes can be identified to explain the outliers - high risk successes and low risk failures? Answers here should address learning objectives. While correlations can be identified by desk analyses,

[58] McGee and Gaventa (2011) IDS Working Paper Volume 2011 No. 383
[59] See B. Befani (2012) Models of Causality and Causal Inference, on the importance of identifying (testable) causal mechanisms to explain associations,
[60] SVG&W question set 20
[61] http://www.mandenews.blogspot.co.uk/2012/06/open-source-evaluation-way-forward.html
[62] http://mandenews.blogspot.co.uk/2010/10/do-we-need-minimal-level-of-failure-mlf.html

analysis of the outliers will require attention to specific projects, not only via their documentation, but also through the use of any scheduled evaluations.

5.2. Refining evaluation questions, and expected answers

Lessons can be learned from other approaches to accumulating knowledge.3ie have commented about the lessons learned from systematic reviews:

"Many of the review questions development researchers have attempted to answer in systematic reviews seem too broad, which inevitably leads to challenges. There is a trade-off between depth and breath, but if our goal is to build a sustainable community of practice around credible, high quality reviews we should be favouring depth of analysis where a trade-off needs to be made.[63]*"*

The alternative to broad open-ended questions is testable claims (hypotheses or predictions). These should be easier to identify after the policy teams are able to identify clusters of projects, each of which is addressing specific kinds of outcomes, rather than thinking about all policy relevant projects as a whole set. The quality of questions is likely to be further improved if they are developed as part of a hypothesis building stage of an evaluation process, one which makes use of prior evaluations and external evidence sources. In this context, the use of scheduled evaluations would be a subsequent part of the overall process rather than the sole activity. The proposal for a hypothesis building stage is discussed in more detail in section 7.

Outside of the world of development projects there are other useful approaches to building knowledge about what works where, which are not dependent on hypotheses formulation and testing, and thus retain some openness to the unexpected. Businesses that accumulate large data sets on their customers' behaviour typically use suites of data mining tools, to identify what combinations of customer attributes are found to be associated with what kinds of purchasing decisions.[64] However, this is done with a *defined* set of attribute and outcome data. This kind of exploration could be done with policy relevant projects, only if and when there is an adequate data base of information about those projects. Then, DFID could be asking questions such as "*With what combination of conditions do girls achieve better secondary school completion rates than boys?*[65]

The process of designing evaluation questions needs to be matched by some attention to the type of explanations they might generate. In Annex J distinctions are made between three types of explanations that might be developed (single factor explanations, compound explanations, and multiple compound explanations). With complex interventions of the kind found in many E&A and SVG&W projects it is likely that there will be few single factor explanations and more need for multiple compound explanations (where more than one set of associated conditions are needed to explain the full set of observed outcomes).If so, the choice of appropriate means of representing explanations will matter. While narrative descriptions will help, the use of Decision Trees will improve their communicability and testability. The use of Decision Trees is explained in Annex H, and could be used to inform the work of those who will be involved in the Next Steps described in section 7.

[63] http://www.3ieimpact.org/userfiles/doc/SR_blog.pdf, accessed in April 2012

[64] Shmueli, G; Patel, N.R., and Bruce, P. C. Bruce (2010) Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner. Wiley.

[65] Where the conditions examined would describe both types of context and interventions.

6. Conclusions

6.1. Readiness for macro-evaluation

Neither the Empowerment and Accountability Policy Area, nor the Strategic Vision for Girls and Women, is ready for a macro-evaluation because:

Basic documentation is not yet available for the majority of currently operational DFID projects (see section 3.1).

Where documentation is available it is not yet clear which projects is 'policy relevant' to each area (see section 2.4).

Addressing the two issues mentioned above is critical to DFID's concern for *accountability.*

6.2. The proposed approach to macro-evaluation

According to the TOR for this evaluability assessment, the proposed approach to the E&A and SVG&W macro-evaluations was to appraise and then synthesise the data available in scheduled evaluations to build generalisations about achievements in each policy area. Where the evaluability assessment identified a need, the process would include the commissioning of specific "thematic, cross-cutting and sectoral studies" and/or a synthesis of other research[66].

The large number and diversity of projects involved makes any attempt to find widely generalisable answers to evaluation questions a big challenge, if not impossible. For that reason, a macro-evaluation that would try to do so is not advised. Instead it would be more practical to focus on clusters of projects with comparable outcomes.

The conclusion of this evaluability assessment is that the reliance on existing or scheduled evaluations as the primary unit of analysis is problematic (see section2.5).The set of evaluations that could be analysed and synthesised is representative of a currently unknown population of projects. This information will become available over time, but is neither very predictable nor controllable. The alternative is to treat projects as the primary unit of analysis, and to seek to understand these through the use of mandatory project documentation and scheduled evaluations. Subject to the availability of project documentation (which affects everything), sets of projects can be deliberately selected and worked with.

Often attempts to syntheses results from evaluations do so through a snapshot approach, by examining a collection of evaluations available at one moment in time and creating generalisations about them. The alternative, which has been foreseen by the TOR, is to take a more dynamic approach and build knowledge over time. Within a cluster of projects with comparable outcomes there is likely to be a sequence of scheduled evaluations that will present opportunities to test both conclusions and newly emerging hypotheses that have been built up beforehand, from project

[66] See page 2 and page 4 of TOR, Annex A.

documents, previous evaluations and external research. This approach requires significant investment in data gathering and analysis activities outside of scheduled evaluations. Evaluations would be seen as one part of a process of knowledge building and testing, not the entire process.

Not all the evaluation questions of interest to DFID policy teams will be addressed by such an approach. Commissioned surveys will be needed for the required policy implementation review of the SVG&W (see section 5.1.1.). Analysis of existing data will be the focus of others, such as the analysis of risks and failure rates (section 5.1.7), and comparative project achievement ratings section 5.1.4). The issue of incomplete project documentation and lack of data will be a problem regardless of what approach is taken to evaluating the two policy areas. A focus simply on synthesising the results of evaluation will not avoid it.

7. Proposed next steps– an iterative "knowledge building" process

In this section, a process is proposed to meet DFID's requirements for macro-evaluations of E&A and SVG&W. The process addresses the evaluation purposes of *accountability* and *learning*. It also meets other key requirements, including that evaluation questions be revised and customised with each scheduled evaluation, and that the policy level ToC be revised periodically.

The proposed process has five key features

1. Priority is given to collation of basic descriptive information, and the setting of boundaries through identification of policy relevant projects, before the use of any evaluations.
2. Policy relevant projects are the primary unit of analysis. Analysis of mandatory project documentation would be used alongside scheduled evaluations to seek answers to evaluation questions.
3. Rather than a snapshot approach of generating lessons, the process is iterative, allowing for a build-up of knowledge over time.
4. Improved policy area ToCs are seen as a product of such the evaluation process, rather than a pre-requisite to commissioning evaluations.
5. There is an annual reporting cycle rather than rather than an interim and final report.

Figure 1 over the page provides an overview of the proposed process. There are three kinds of expected *outputs* described in the model:

1. The first is a description of the portfolio of policy relevant projects. The public availability of this information alone can be an important form of accountability, in addition to being an essential basis for subsequent planning and evaluation activities.
2. The second is a proposed annual synthesis report, using knowledge from recent evaluations, analysis of available project data, and external evidence sources. Reporting should be coordinated with the DFID corporate reporting cycle
3. The third would be an update of each policy area ToC and its associated evidence base.

The proposed process has six key activities needed to reach these outputs; some of these activities will need to be done internally by DFID and some could be contracted out.

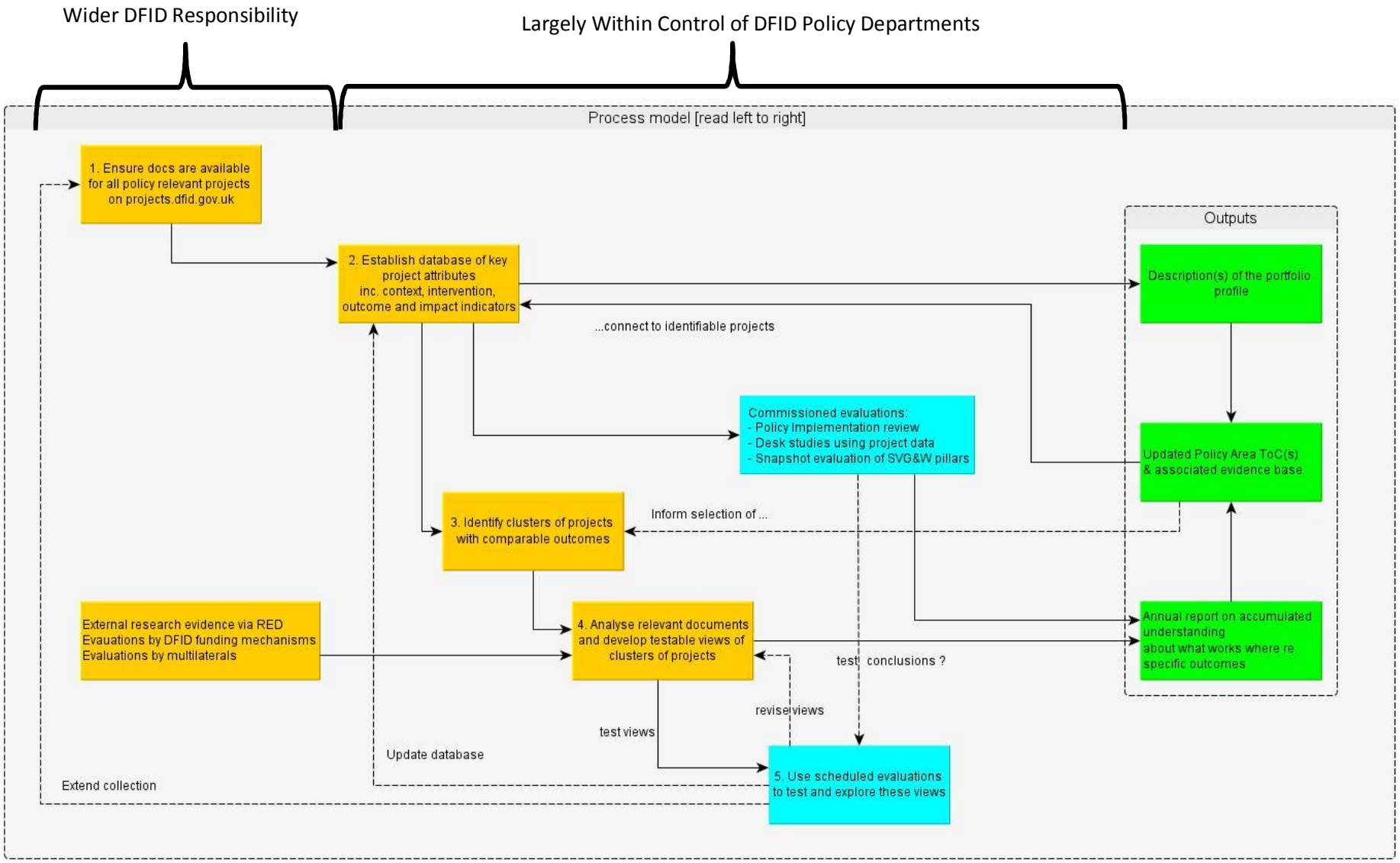


Figure 1: Diagram of Proposed Next Steps

Step 1: Improve the coverage of project documentation

As explained in section 3.1, access to adequate documentation is a bedrock issue. At present key documents, such as Business Cases and LogFrames, are only available for 40% of post 2010 projects. It is not possible to build a comprehensive description, let alone do an adequate evaluation, of DFID's work in the two policy areas until this problem is addressed.

DFID Policy Division and EvD staff need to explore all possible avenues for increasing the document coverage of post-2010 projects. The focus should be on increasing the coverage of the projects in the 28 focus countries. Completion of the *projects.dfid.gov.uk* website is not a task that could be outsourced as part of a contract for evaluation services. It is an on-going internal process, over which Policy Division and EvD staff have limited influence.

Making progress with document coverage should be of wider interest within DFID and beyond. With the *projects.dfid.gov.uk* website, DFID is in effect encouraging what could be called "open source" evaluation, where anyone with access to the basic information can start to develop their own analyses of DFID's work.

Step 2: Build a database of policy relevant projects in operation post-2010

A single database (covering SVG&W and E&A) of policy relevant projects is needed for two reasons: to produce a descriptive profile of all DFID investments in SV&GW and E&A projects (for accountability) and to enable a filtering of projects (to identify policy relevant projects, and clusters of projects with common outcome measures).

The following kinds of information need to be available via such a database:

1. Project numbers, which can be cut and pasted from Excel files downloadable from *projects.dfid.gov.uk*
2. Context descriptions, available within Business Cases, but requiring human judgement
3. Intervention descriptions, mainly available within Business Cases, and also human judgement
4. Target group descriptions, available in both LogFrames and Business Cases
5. Impact and outcome descriptions, which can be cut and pasted from the LogFrames available online. If available at project level, data on "we wills" measures should also be included
6. User–defined tags, which enable as-needed classification and filtering of projects, using any fields in any of the 1-5 categories above.
7. Where possible, hypertext links to documents sources at *projects.dfid.gov.uk*, for 1-4 above

The volume of work involved needs to be recognised. It is estimated that approximately 200 new projects came online from the 28 focus countries in 2011, and additional (though lower) numbers are expected in subsequent years. All of these would need a policy relevance ranking and then an estimated 47% (94) would need detailed database records of the kind suggested above. An annual updating of existing policy relevant projects may also be required, e.g. hypertext links to Annual Reviews and Project Completion Reports.

Given the on-going nature of this work, and the need for close and frequent engagement with DFID staff, much of this work would ideally be carried out in-house. However, if a new job cannot be created to complete these tasks, or current job-descriptions re-designed to that effect, much of the

work *could* be contracted out. While policy relevance ratings could be outsourced, it would be essential that there is a process for validating these, involving random checks by DFID staff in country offices and/or Policy Division.

Step 3: Identifying clusters of policy relevant projects

The proposed database would be used to identify clusters of projects with comparable outcome measures, to facilitate explanations about what works in which context. Ways of doing this have been outlined in section 3.2 of this report. The membership of the clusters of projects will also change over time, with completed projects dropping out and newly established projects included. This work could be contracted out. However, given the variety of possible groupings and the number of possible clusters of projects, the process would need to be directed by the priority interests of the two policy teams, which would involve substantial DFID engagement with the contractor.

Step 4: Developing testable views of projects within clusters of comparable projects

The development of hypotheses to be tested with each scheduled evaluation will require time and expertise and will need to be contracted out.

The focus should be on candidate causes of:

- Differences in performance of the projects in the cluster due to types of interventions (E&A question set 2), including (but not limited to):
    - Interaction effects of combined interventions (see section 5.1.2).
    - Mainstreaming versus specialising of interventions (see section 5.1.3).
- Differences in performance of the projects in the cluster due to types of contexts (E&A question set 5, and SVG&W team's question set 12) (See section 5.1.6).
- Projects found as outliers in terms of risks and failure relationships (see section 5.1.7).
- Projects which are outliers in VfM terms (see section 5.1.5).

Sources could include:

- Analysis of existing data available within ARIES and policy area databases (mentioned above) e.g. costs data, risk ratings, achievement ratings, and coded attributes of interventions and contexts.
- Analysis of narrative contents of project documents (Business Cases, Annual Reviews and Theory of Change).
- Analysis of policy area ToC (diagrammatic and narrative versions)
- Findings from previous evaluations of related DFID projects, before 2011.
- Findings from evaluations by DFID's centrally managed funding mechanisms[67]
- Findings from evaluations carried out by other aid agencies
- Findings from relevant research, as identified by RED or others.

The results of this step will need to be documented before the next step is undertaken. Prior declaration of hypotheses to be investigated is a recognised means of prevent under reporting of negative findings (both by researchers and by publishers) and is now recommended good practice.

[67] E.g. Programme Partnership Arrangements (PPAs), the Global Poverty Action Fund (GPAF) the Governance and Transparency Fund (GTF), or the Girls Education Challenge Fund.

Step 5: Using scheduled evaluations to test and analyse views

Recent literature on the analysis of impact and causal attribution suggests that there two necessary elements to valid claims[68]. The first is evidence of *co-variance or association*. For example, between cases of risk and failure or between a specific aspect of the context and a particular kind of project outcome. This is the kind of data that can be extracted from a functioning database on policy relevant projects, and more specifically, from data on clusters of projects with comparable outcomes. The second is a testable claim about *mechanisms*, the ways in which one event is expected to affect another event with which it is correlated. The more detailed and observable the proposed mechanism is, the more testable it is. Case studies are opportunities for testing if such mechanisms are at work or not. A sequence of evaluations within a cluster will provide a sequence of case study opportunities. The same evaluations will also be important opportunities for checking the reliability and validity of the project data held in the policy area database, especially those which make up parts of significant associations.

At Step 4 it will be necessary to identify both associations and mechanism to be tested. E.g. which project is a VfM outlier and why it has become so.

It is not clear what proportion of projects in a cluster are likely to have scheduled evaluations. Of the 292 post-2010 projects listed on the *projects.dfid.gov.uk* in May 2012, a third may be policy relevant projects (say 100). Nineteen of these are known to have scheduled evaluations, or approximately one in five. More may be planned as projects get nearer their completion dates. Wider coverage of evaluations may be needed to ensure sufficient opportunities for building knowledge about each cluster.

Testing of hypotheses about what is working and why will require active liaison with those responsible for planning and managing scheduled evaluations. The purpose of this liaison would be to find out what sort of data will be collected, how and from whom, and where necessary to request additional enquiries. Cooperation might be more available if something can be offered in return, such as access to the policy area database proposed above, which would provide wider comparative data.

The process of using scheduled evaluations in this way would need to be contracted out, possibly to the same party responsible for the previous activities (activities 2-4 listed above). They would not be responsible for the scheduled evaluations but would be responsible for negotiation with planners of scheduled evaluations, to have their data requests and evaluation questions included in the evaluation plan.

Step 6: Using commissioned evaluations for special purposes

In section 5 above, it was noted that there were some evaluation questions that would need specialised evaluations or additional work, rather than simply making use of scheduled evaluations. These include:

[68]Befani, B. (2012) *Models of Causality and Causal Inference*, An annex to Broadening The Range Of Designs And Methods For Impact Evaluations, by Elliot Stern et al; Cummins, D. (2012) *Good Thinking*. Cambridge University Press (see Chapter 6)

- A policy implementation review (of the SVG&W), using document reviews and staff interviews (section 5.1.1).
- Analysis of overall achievements, using PCR achievement ratings (section 5.1.4)

The evaluability assessment team were asked to comment on the extent to which the proposed macro-evaluation of the whole Strategic Vision would meet the requirement for "pillar evaluations". The evaluability assessment conclusion is that it may not be necessary to commission separate pillar evaluations in addition to the process described in this report. All four Pillars should benefit from progress with step 1, above (document coverage) and with the implementation of steps 2 and 3 (database development and identification of project clusters). The suitability of the sequential approach to assessing clusters of projects under each pillar will depend on: (a) the deadline for each pillar evaluation and what is known about the timing of scheduled evaluations in their relevant clusters. Where there is least conflict in timing, some additional evaluations may need to be encouraged or even commissioned. If timing requirements are in serious conflict then a sequential approach may need to be abandoned. The alternative would be a snapshot approach, through a commissioned evaluation that looks at multiple projects in relatively little depth.

Key outputs

A major output of the whole process would be two annual syntheses reports, one for each policy area. These would summarise the current explanations for the variations in outcomes in each of the clusters of projects being investigated, including the evidence accumulated during the year from use of the scheduled evaluations and any available external sources.

The commissioned evaluations would need their own free-standing reports, but should also be summarised in the annual synthesis reports.

The development of the proposed database will enable the production of another output, a descriptive profile of the "portfolios" of projects covering the two policy areas. If the database is properly designed it should be possible for DFID policy teams to access this database and develop their own profiles, as needed. However, an annual version of this task could be contracted out as one of the expected sections of the annual synthesis report.

Description of the portfolios should include some analysis of what are not there, the contexts not covered by the projects and the interventions not underway. This would need to be informed by access to evaluations of the same policy areas by other agencies, and of projects funded through other mechanisms within DFID.

There will be checks and balances on data quality and validity of the analyses. DFID now has a policy of transparency in respect to all evaluation reports, which would include the above. Scheduled evaluations will provide opportunities to check the accuracy of data in databases, especially attributes ascribed to projects, but also data collated from project documentation. The analyses developed then tested through scheduled evaluations will be open to further testing as newly established projects that get added to project clusters. Can they accurately predict the types of performance that will be seen in the new projects[69]?

[69] This is the standard way in which results of data mining analyses are assessed, by taking results developed with one data set (known as training cases) and applying them to another (known as the test cases)

Developing the policy level Theories of Change

As noted in section 4 of this report, the ToCs of the E&A and SVG&W policy areas should be seen as part of an effort to accumulate a body of evidence based and testable knowledge about what works in what circumstances to deliver E&A /SVG&W results. In other words, a *product* more than a *pre-requisite* of macro-evaluations. That is why, in the process model above, the ToC is the last step in the sequence, albeit one with feedback links to earlier stages.

Although the focus of discussion in section 4 has been on the diagrammatic summaries of this knowledge, the body of knowledge being accumulated by each policy area would obviously need to be more extensively documented in narrative form. The diagrammatic ToC provides a usable overview of the different pathways to achieving the policy goals.

The approach outlined here would require an annual updating of the ToC, after receiving annual synthesis reports mentioned above. How much time is invested in doing this could change year from year, and be up to the policy teams. Development of the ToC could be assisted by an external facilitator with relevant expertise.

This annual 'health-check' would allow current evaluability problems with the ToC to be addressed. One being the lack of identifiable causal linkages between different events in the ToC. This is especially the case with some of the pillar ToC (see section 4). The second problem is how to verify whether the events in the ToC are taking place. At the outcome/impact level this can be addressed by clear statements of relevant indicators. At the earlier and middle stages of the ToC the solution mentioned above was to find exemplar projects that illustrate the changes described by various linkages. Other improvements would be the identification of any elements within the ToC which are found to be necessary or sufficient causes of the outcomes.

The updating of the policy area ToC should be informed by evidence from other sources in addition to the scheduled and commissioned evaluations discussed above, including evaluations by other aid agencies and macro-evaluations of activities funded by other centrally managed mechanisms within DFID.

Management arrangements, time-frames, and budget

Management arrangements

Given the high level of overlap between investments of policy relevance to both the E&A policy area and the Strategic Vision for Girls and Women, it is recommended that work commissioned to address the data issues raised above, and to prepare for subsequent evaluation work, be commissioned jointly by the two policy teams.

Not all the steps listed above necessarily fall into one contract or piece of work. The steps include some tasks that are deemed essential elements for developing a more evidence-based policy making and programming culture in the E&A and SVG&W teams (steps 2-5), and others (mentioned under step 6) which respond to specific areas of interest to DFID but which are discrete one-off evaluation exercises. However they would ideally be commissioned after steps 2 and 3 (establishment of the database and analysis of clusters of projects) have been completed.

It is important to note that while many of the tasks could potentially be contracted out, they will all require substantial on-going engagement with DFID staff.

Time frames

Steps 2-5 have an initial 'establishment' phase that could be contracted and delivered within a set time frame. Accessing documents associated with post-2010 projects and filtering them for policy relevance is likely to take at least five months. Developing a database that includes detailed information on all filtered (policy relevant) projects would take an additional three months, possibly concurrent with the policy relevance rating work.

Once the systems are established, they will require constant updating, periodic interrogation, and regular use by the contractors and by DFID staff.

Step 3 (identifying clusters of projects with comparable outcomes) should not be as time-consuming, provided the database has been done. But additional time would need to be committed by DFID staff to prioritise those clusters that should then be the focus of in-depth analysis (step 4). Because the number of clusters of projects with comparable outcomes is unknown, it is hard to conjecture at this stage how long step 4 would take. Four weeks per cluster may be needed, given the importance of sourcing evidence from outside DFID as well as from within (both project documents and all relevant evaluations). Ideally there should be sufficient time to make use of all evaluations scheduled for 2013.

Planning of work in subsequent steps will not be possible until step 3 is completed, and even at that point will only be tentative. This is because the timing of many of the evaluations planned for policy relevant projects will still be uncertain.

Budget

The evaluability assessment team has developed some very provisional costing for carrying out the proposed steps listed in section 7 above. Assuming steps 2, 3, 4 and 5 were commissioned as one contract, with the same contractor being responsible for the annual reports and the annual facilitated update of the ToCs, an estimated budget of £100,000 per financial year would be required. Note that there would be high start-up costs, such that, even if work does not get commissioned until November 2012, the budget for the 2012/2012 financial year will be similar to that needed for subsequent years.

The costs of the 'policy implementation review' of the Strategic Vision for Girls and Women will depend on the extent and nature of the questions it seeks to answer; a focussed, relatively light-touch review might require £50,000.

8. Annexes

8.1. Annex A: Terms of Reference

Evaluability Assessments for Evaluations of 'Empowerment and Accountability' and DFID's Strategic Vision for Girls and Women

These terms of reference cover work to be contracted by DFID to undertake two evaluability assessments as part of a wider design processes for macro-evaluations of two important DFID policy areas: empowerment and accountability (E&A) and the DFID Strategic Vision for Girls and Women.

These TOR should be read in conjunction with the concept notes that have been developed for both macro-evaluations.

A. Background

DFID has undertaken to evaluate the impact of and approach to implementation of two key policy areas, both agreed in 2011. Developing and implementing both of these macro-evaluations will be a challenging process, with multiple components, stakeholders and deliverables. DFID has opted to undertake the preparation and design work in a number of stages, to allow for reflection and consideration of different options. There are clear areas of overlap between the broad theories of change underpinning the two policy areas, and the programme areas that the two evaluations are likely to look at. DFID is therefore contracting a single team to undertake evaluability assessments for the macro-evaluations of both E&A and girls and women. Once the assessments are concluded, we will determine whether it is desirable and feasible to continue working on the two evaluations jointly or whether separate programmes will be more appropriate.

Empowerment and Accountability

In February 2011, DFID's Development Policy Committee endorsed a proposal that DFID should do more to enable poor people to exercise greater choice and control over their own development and to hold decision-makers to account. Our conceptual framework for this includes a number of linkages between donor supported interventions that seek to enable different forms of empowerment (economic, social, or political) and accountability, in the expectation that improvements in empowerment and accountability will deliver better development and growth outcomes for the poorest. A schematic summarising our current policy model provides an overall idea of how E&A works (Annex 1) but this will be an 'evolving object' between now and 2015/16. DFID's focus on empowerment and accountability will be implemented through a range of programmes designed and implemented at country level, either as interventions with core objectives on E&A, or as components of broader programmes in particular sectors.

Strategic Vision for Girls and Women

The UK has put the empowerment of girls and women at the heart of international development. DFID's Strategic Vision for Girls and Women, launched in March 2011, identifies four priority pillars for action to deliver real change for girls and women:

Pillar 1: Delay first pregnancy and support safe childbirth
Pillar 2: Get economic assets directly to girls and women
Pillar 3: Get girls through secondary school
Pillar 4: Prevent violence against girls and women

Achieving results across these 4 pillars also depends on improvements in the enabling environment – i.e. the attitudes, behaviours, social norms, statutory and customary laws and policies which constrain the lives of adolescent girls and women, and perpetuate their exclusion and poverty.

The Strategic Vision has wide ranging implications for DFID and is being implemented through a large number of programmes developed across DFID – by country offices, Policy and Research Division, Private Sector Department, Civil Society Department and International Financial Institutions Department. The evaluation could therefore be very wide ranging in scope and its parameters will need to be carefully determined in the design stage. It is likely to take an interest in some or all of the following:

- Overall impact of the Vision
- Impact of each individual pillar of the Vision
- How work on the enabling environment has been taken forward
- Interaction between the pillars, and between the pillars and enabling environment work
- Institutional arrangements for developing and driving the Vision

The evaluation will need to draw on other evaluations – it will be in large part dependent on the findings of evaluation activity commissioned by other DFID business units, including country offices and Vision pillar leads. It is possible that additional evaluation work might need to be commissioned, for example to assess the impact of the enabling environment. The assessment should make recommendations on this possible requirement.

DFID's approach to 'macro-evaluation'

We are using the term "macro-evaluation" to incorporate the two key concepts:

- Synthesis: bringing together the findings of existing evaluations which have used various methods.
- Meta-evaluation: reviewing evaluation methods – rather than content – to validate the quality of material

For the macro-evaluations, there are two key objectives (or purposes) that the evaluation designs will need to support:

- Accountability: what has been achieved from investments in different countries (e.g. what have been the effects from DFID-funded interventions?)
- Learning and evidence: what works and what doesn't, and how does context matter. What generalisable questions can be drawn?

The macro-evaluations will involve synthesising findings and lessons from evaluations conducted mainly by devolved offices. They will require:

- A minimal 'common evaluation framework' so as to allow for comparison and analysis of common themes
- A sufficient number of 'component' evaluations following this framework to synthesise across settings, themes, ways of implementing and in different combinations – this implies some leverage/negotiation etc. with devolved offices to ensure that they help meet evaluation needs.
- A quality assurance process that ensures that the component evaluations are of good enough quality (reliable, valid, defensible, well-conducted etc.) to synthesise them.

DFID's approach to evaluability assessments

DFID's use of the term "evaluability assessment" goes beyond examination of a programme's coherence and logic from an evaluation perspective. We expect the assessments will clarify evaluation questions; assess complexity and evaluability concerns; assess relevant evaluation approaches; consider budget implications; assess availability of data sources; and set out timeframes and milestones.

B. Scope of work for Evaluability Assessments for Empowerment and Accountability and DFID's Vision for Girls and Women

1. **Evaluation Questions.** Define and recommend core evaluation questions for each evaluation. A suggested list of possible questions for each macro-evaluation is at Annex 1.

2. **Programme attributes.** Define the unit(s) of analysis for each evaluation and draw boundaries around the scope of programmes to be included. Define and map programme attributes; highlight implications for evaluation designs and assess to what extent the evaluations will ensure coverage across programme types and attributes.

3. **Data availability.** Examine existing data sources, including planned evaluations, and assess whether data generated is likely to meet macro-evaluation needs. Advise on whether additional data will be needed to enable comparison / generalisation and whether it will be built into evaluation design, e.g. thematic, cross-cutting and sectoral studies; or syntheses from wider research.

4. **Theories of Change**. Assess existing theories of change, and provide recommendations for linkages and improvements, and in consultation with DFID, revised change models for the macro-evaluation. Examine opportunities for an 'evolving' ToC for the macro-evaluation, and suggest at what points and how ToC and EQs could / should be revised.

5. **Timeframes.** Suggest an appropriate timeframe with key milestones for reporting over the 3 year period for both evaluations. It is expected that there will be at least 2 interim reports during the period. Provide suggested timeframe for reviewing and revising ToC and evaluation questions.

6. **Budget.** Provide an indicative budget for each evaluation, and for each financial year from 2012 to 2015/16

7. **Management issues:** Assess and recommend whether the two evaluations can be managed jointly, or whether it would be more feasible to run as separate processes. Recommendations in response to issues set out in section C (below) should be included.

8. **Terms of reference** should be prepared for each evaluation.

C. Evaluation challenges - Empowerment and Accountability & the Strategic Vision for Girls and Women

There are a number of complicated challenges that are common to both evaluations, which the evaluability assessments should consider within the scope of work set out above.

Managing a complicated evaluation process

Identify the most appropriate approach and mechanisms for meeting the challenges involved in developing and implementing evaluations of E&A and Vision for Girls and Women, considering:

- The complicated nature of both E&A and Girls and Women as policy areas;
- The need to establish boundaries around what policy space each evaluation will cover (e.g. with E&A, should it include work on political parties, elections, political empowerment; with the Vision for Girls and Women, whether to include work on each of the four pillars);
- Most implementation will be done through country offices rather than led by HQ, though the latter is responsible for the macro-evaluation. There may be institutional structures and capacity issues to address;
- Multiple and wide-ranging sets of stakeholders that need to be brought into the process; and
- Overlaps with other policy areas (e.g. service delivery work, financial inclusion, social development, security and justice).

Identifying the programme matrix (population, attributes)

Given the breadth of both policy areas, the evaluations are likely to include several different programme approaches:

a) Single 'dedicated' programmes e.g. improving access to economic support for women
b) Linked sets of programmes – economic, social, political – i.e. packages of interventions?
c) Series of similar programmes (similar goals, target groups etc.) implemented in different contexts
d) Other initiatives that include components or objectives on E&A and/or Girls and Women.

The assessments should recommend an appropriate mix of programmes (existing or planned) to be the basis of the macro-evaluations and highlight relevant implications for evaluation designs.

We have identified a 'long list' of programmes being implemented through DFID country offices that include potentially useful evaluations for these macro-evaluations. The evaluability assessment should review available documentation on those programme / evaluations, and determine whether they are suitable for the macro-evaluation.

Whether supporting work is needed

The Concept Notes outline other possible inputs to the evaluation, and the evaluability assessments should assess and inform to what extent (and if so, how) the macro-evaluation should include these aspects:

Thematic, cross-cutting and sectoral studies
The assessment should identify additional work that may be needed in this area before an evaluation framework can be developed. The Evaluation Framework will be designed in the next design phase of the macro-evaluation.

Synthesis of wider research
To what extent a synthesis of wider research should be included within the macro-evaluation.

Programme level theories of change
In addition to the macro-level 'change models' (Annex 1), each programme implemented through DFID country offices will have a dedicated theory of change. The evaluability assessment should 'scan' those theories of change to identify areas of commonality, difference, and likely changes over time, and how these relate to the macro-level change model.

Given the highly innovative nature of some E&A and programming on girls and women, a balance will need to be struck between a top down design, preset questions, start-up Theories of Change etc and being open to completely different understandings that might come from the bottom up experience of actual evaluations. The assessments should examine opportunities for an 'evolving' ToC, whilst recognising that there will need to be some clear outputs (for accountability purposes) and specific lessons (to ensure relevant learning and evidence is generated). The assessments should suggest at what points in the evaluation process, and through what means the starting ideas (Evaluation Questions, theories of change etc) will be revised.

Timing and Phasing
It is highly unlikely that a set of evaluations will be commissioned at one point in time across the country programmes on which the evaluation will be based. This offers opportunities (some of which may be challenging). For example results or interim outputs from one evaluation could feed-in to the conduct/design of latter ones. Problem x has come up in one place, which might lead to this being specifically examined in another programme.

This also creates opportunities for 'networking' and mutual learning over the next 4 or 5 years. Involving programme managers and stakeholders in learning how to implement and specify evaluations better – and making explicit their tacit knowledge creates opportunities which evaluability assessment should identify and recommend how to maximise these opportunities.

The non-standard time-line of this work together with the nature of E&A and Girls and Women (often bottom-up, requiring commitment and participation of different stakeholders etc.) also creates opportunities for participatory and action research approaches. For example this might involve agreeing with local stakeholders that there are two strategies they might want to try and then comparing the results; or involving stakeholders in interpreting results before conclusions are drawn.

D. Outputs
The following outputs should be delivered under this TOR:

1. Two separate reports (for E&A and Vision for Girls and Women) covering points 1-7 in the Scope (section B above):

- Assessment and recommendations on key design and management challenges, including proposals for timeframes, reporting requirements, governance structures, and possible areas of additional work

- A definitive assessment and recommendations on evaluable questions

- Recommendations on inclusion of specific country programme evaluations in the macro-evaluation, highlighting potential challenges, gaps and risks, with mitigation strategies.

- An assessment of existing theories of change, recommendations for better linkages and improvements, and in consultation with DFID, a revised change model for the macro-evaluations.

- Indicative budgets

The following outputs should be completed once DFID has responded to the evaluability assessment reports:

2. TOR for the design and implementation of the macro-evaluation for E&A
3. TOR for the design and implementation of the macro-evaluation of DFID's Strategic Vision for Girls and Women

E. Timeframe & reporting

The final evaluability reports should be submitted by 01 May 2012. Deadlines for interim drafts will be agreed after contracting. The TOR for both evaluations will be due by mid-June 2012.

The contracted team will report to the joint group overseeing the two evaluations:

- E&A: Lu Ecclestone (Policy Division) and Lina Payne (Evaluation Dept.)
- Girls and Women: Rebecca Trafford-Roberts / Teresa Durand (Policy Division) and Zoe Stephenson (Evaluation Dept.)

The contract will be managed by the Politics, State and Society Team in Policy Division (John Howarth).

F. Skills and expertise

The team undertaking this work will need to demonstrate significant experience in the following areas:

Essential

- Experience of designing and implementing evaluations of complex programmes

- Experience of conducting evaluability assessments, or similar, including attention to financial aspects and theory of change
- Understanding and experience of a range of evaluation methods (quantitative and qualitative), and applying quality standards to their use

Desirable

- Experience in the use of relevant standards and norms.
- Familiar with relevant codes of conduct and ethics.
- Familiarity with DFID's approach to programme delivery and management
- Familiarity with / demonstration of Paris Declaration principles (and the Accra Agenda for Action)
- Competence in gender, diversity and poverty analysis
- Experience of gender and/or empowerment and accountability programmes

Annex 1: Possible evaluation questions

Empowerment and Accountability

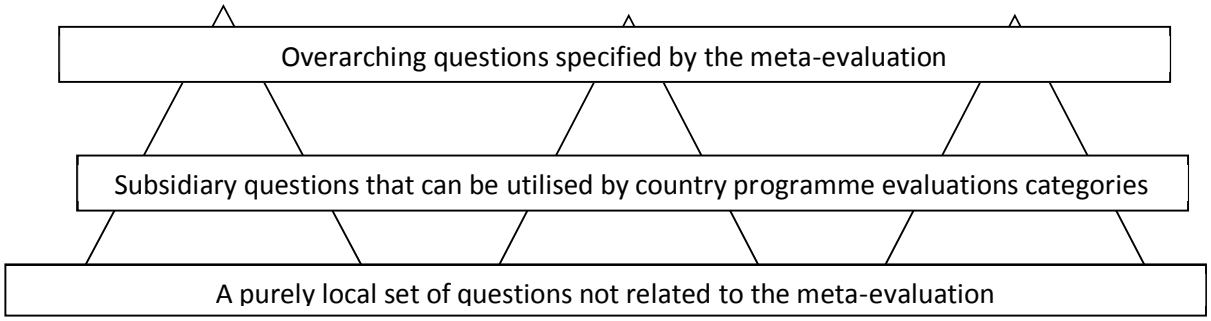
This section summarises thinking to date on how to approach the evaluation questions for the E&A macro-evaluation. The subsequent section does the same for the Vision for Girls and Women evaluation.

Defining the evaluation questions

On-going work to gather and review evidence related to E&A and girls and women, and the associated development of change models have suggested the following approach to evaluation questions – this is neither comprehensive nor prescriptive.

The evaluation is likely to take a layered approach to identifying and responding to evaluation questions. Some will relate to the over-arching policy level, while others will be more specific to particular country programmes and their objectives. The challenge will be in linking the two in a robust manner so that our overall conclusions are rooted in the experiences of country-level programmes.

This diagram illustrates how this might be achieved:



The macro-evaluation will have to balance the need to maintain a consistent sent of questions and areas of focus, whilst allowing for additional areas to emerge as our learning and understanding deepens. Total clarity is (probably) not possible at the outset!

The over-arching evaluation questions are likely to focus on different aspects of change, for example:

- The overall impact of increased work on E&A: e.g. have efforts on E&A made any difference to poverty, development, fragility, governance outcomes, or to social cohesion or power relations? [comment: we know this question is too broad!]

- Impact of specific interventions in different contexts (what works and what doesn't and why?) Why have some programmes worked well for some groups but not others? how well have programmes been adapted to meet different local contexts?

- Interaction between different programmes working on E&A. For example, does working on a number of areas within the policy frame lead to better results than working on individual programmes – what is the sum of the parts? Does empowerment in one sphere (e.g. economic) lead to accountability in others (e.g. political) or vice versa?) Do interventions that combine *both* empowerment and accountability elements achieve better development results than those working solely on empowerment or solely on accountability?

- Institutional arrangements for supporting E&A work: e.g. is it more effective to support 'stand-alone' voice and accountability interventions, or to integrate E&A into other programmes?

- Mid-level strategies on E&A: e.g. What should be the priorities in different contexts? Do some strategies work better in certain contexts and can we generalise about these contexts (build typologies)?

- Drawing out implications for the future: for example, what can we do to improve the success of interventions now underway? How can this intervention be scaled-up or diffused to other settings?

These questions relate (broadly) to the outcome, impact and super-impact levels of the E&A change model (see Annex 1). We expect that it will be challenging to identify clear attributable impact to

DFID interventions, but we hope that linkages to changes at the process / output / outcome level will be demonstrable.

These overarching questions will need to be elaborated in greater detail at central and country levels, as part of this TOR. They will also need to be validated by stakeholders. Discussing these questions will also serve an awareness-raising/educational function across the organisation, and particularly for those offices that will be involved in the evaluation.

Country programme questions

Most country programmes are likely to focus on discrete areas of the E&A change model, and so there will be a set of questions related to these levels of change. The overarching questions for the meta-evaluation will to relate to those through a set of optional questions that country might want to adapt, and questions that may have nothing to do with the overarching questions but would be chosen by offices to meet their own needs. An initial set of questions could include:

- Whether interventions to support greater transparency and accountability have led to more accountability / demand for accountability / reduced corruption?
- Whether programmes have impacted the linkages between citizens and states / officials / service providers, including democratic institutions and processes?
- Whether responsiveness to poor people from state bodies / service providers has increased?
- If there is increased choice and access to services for people, and in which groups (and why)?
- The role of tools such as media and new technology in empowering people and groups?
- What changes in women's social and political participation result from interventions?
- Whether the poor are more engaged in coalitions to achieve policy change or expanded political settlements?

A key objective of the evaluability assessment is to define an appropriate set of evaluation questions.

DFID's Strategic Vision for Girls and Women

The evaluation questions will be determined in consultation with key stakeholders during the evaluation design phase, and will address the DAC evaluation criteria of effectiveness, efficiency, relevance, impact and sustainability. Specific questions may include:

i) Questions on Overall Strategy

- Were the four pillars the most strategic and effective entry points through which to promote girls and women' empowerment? *(DAC criteria: relevance)*
- What impact has the Strategic Vision had on DFID's relationship with external partners?*(relevance)*
- To what extent was the Strategic Vision design coherent, logical and innovative (*Effectiveness)*
- What effect did the Vision's particular focus on girls have on DFID programmes? (*Effectiveness)*
- Has Value for Money been achieved in delivering the vision?

ii) Questions on Results

- To what extent are the results achieved by the vision quantifiable and measurable?

- Did the Vision achieve the results it set out to? (*Impact? Effectiveness?)*
- What impacts has the Vision had on empowerment of girls and women? (*Impact*)
- Was value for money achieved in delivering these results, at all stages of the results chain?
- Do results depend on interaction between pillars? What difference does it make to girls and women if they benefit from more than one area of the Vision? *(Impact and Effectiveness?)*
- Do impacts depend on progress in the enabling environment? (*Impact)*
- To what extent are the results achieved to date and future results likely to endure into the longer term? *(Sustainability)*

iii) Questions Implementation of the Vision

- How has the Vision guided the work of DFID Country Offices on girls and women? (*Effectiveness)*
- How has implementation varied across the organisation? *(Effectiveness and Efficiency)*
- Are there effective cross-pillar linkages? Are there effective linkages between the four pillars and the enabling environment? If so how have these been achieved? *(Effectiveness and Efficiency)*
- Have different approaches to implementation affected the extent to which value for money has been achieved*?*

iv) Questions on the institutional arrangements

- Do the organisational structures for the Strategic Vision provide clear leadership, a strong accountability structure and positive incentives for effective delivery of DFID's work on girls and women? *(Effectiveness and Efficiency)*
- Has the Vision led to an increase in the allocation of financial resources to programmes on girls and women? And increased mainstreaming of girls and women in DFID programmes? If so, has our ability to track spending on girls and women changed as a result of the vision? *(Effectiveness)*
- How have reporting requirements in the Corporate Performance Framework affected implementation of the Vision? *(Effectiveness)*
- How effectively did DFID respond to risks identified in the Vision and to changes (opportunities and challenges) in the external environment? (*Effectiveness)*

More specific questions on the pillars will be determined by pillar leads.

8.2. Annex B: Methodology

The methodology for this evaluability assessment was experimental and iterative. There is no blue print for conducting evaluability assessments of this scale. The section below describes the specific tasks undertaken and relates them to the Scope of Work outlined in the evaluability assessment TOR.

Meetings, interviews, and workshops with key stakeholders

On 27 and 28 March 2012, the evaluability assessment team met with the DFID E&A and SVG&W teams for two inception meetings, each of two and a half hours, with a view to clarifying the Terms of Reference and discussing a proposed approach to the work. On 26 and 27 April 2012 the team held workshops with each policy team to ***discuss the policy level Theory of Change***, and to undertake ***a Hierarchical Card Sorting exercise*** (described in Annex E) with a view to ***identifying a sample*** of seven priority countries per policy area to be included in the macro-evaluation, and ***exploring DFID's hypotheses about the specific contextual factors*** that make a difference to project implementation and outcomes. The results of the card sorting exercise for each team are presented in Annex F.

In addition to these meetings, interviews either in person or over the phone were conducted with a number of DFID staff, as well as two members of the PPA E&A learning group, with a view to ***exploring priority evaluation questions***. There was also regular email and phone communication between the evaluability assessment team and core DFID staff to clarify a number of issues as they arose during the course of the evaluability assessment. A list of individuals consulted as part of the evaluability assessment is listed in Annex C.

Background document review

The team read a wide array of literature around methodologies for evaluability assessments and for evaluating complex programmes, as well as a large body of literature provided by DFID surrounding the background, policy directions, and future evaluation and research plans for both policy areas References for particular evidence sources cited in the document are cited in the footnotes and provided in detail in Annex D, which also contains a short bibliography of evaluability assessment literature.

Review of DFID Website

With the objective of getting key information regarding programme spend, numbers and spread of DFID programmes, and access to key documents, the evaluability assessment team searched the DFID website (www.dfid.gov.uk/About-us/How-we-measure-progress/Aid-Statistics/Statistics-on-International-Development-2011/Key-Statistics), and Programme documents database (*projects.dfid.gov.uk*).

Key Word Search and Policy Relevance Ratings

Part of the work expected of the evaluability assessment was to *"draw boundaries around the scope of programmes to be included"* within the proposed macro-evaluations. In the absence of DFID internal categorisations of projects, such as PIMS markers, to identify which are directly relevant to each policy area under examination, the evaluability assessment team need to develop an approach for identifying policy relevance.

The evaluability assessment team tried two approaches to this. The first was the use of ***key word searches*** of Project titles, Business Cases, LogFrames and Annual Reviews[70]. The second was developing a ***Policy Relevance rating scale*** which was then used by DFID staff and the evaluability assessment team to form judgement of the policy relevance of key projects through a review of the Business Case and/or LogFrame.

The rating scale allowed the use of "fuzzy" categories, which allow partial membership of a category, rather than binary judgements, because it was recognised that in many cases E&A or SVG&W objectives would be only part of a project design. The scale was as follows:

4	The project is wholly focused on policy area objectives
3	The project is largely focused on policy area objectives
2	The project is partly focused on policy area objectives
1	The project is not addressing any of policy area objectives
0	It is not possible to say

The scale was then applied by a DFID staff member from each policy team and by the evaluability assessment team members in parallel.

Key word searches generated unreliable results, not fitting well with policy relevance judgements made by either DFID staff or the evaluability assessment team[71]. DFID and evaluability assessment team judgements about policy relevance using the policy relevance scoring had a high level of agreement, but with DFID staff tending to give higher relevance ratings than the evaluability assessment consultants. Ultimately it is DFID judgements that matter, since DFID staff are responsible for defining policy in the two areas. Ideally it would be DFID staff at the country level e.g. Social Development Advisers (SDAs), who would carry out the policy relevance ratings, because they would be more informed about the designs of the projects under review.

Review of project documents and creation of project database

The evaluability assessment team accessed (via the website, or directly from our clients), core project documents associated with country programme expenditure in seven[72]of the 28 core DFID country programmes.

For these seven countries, the evaluability assessment team reviewed the available Business Cases, LogFrames and Annual Reviews for every project with a start date of January 2011 or later[73].

In addition, the evaluability assessment team looked at documents associated with nine projects with a start date prior to January 2011, identified by the E&A and SVG&W teams as being projects of particular relevance to one or other policy areas.

[70] The words searched for were *empowerment*, *accountability*, *gender*, *girls*, and *women*

[71] A wider range of key words found more relevant projects, but also found many more that were not relevant

[72] Ethiopia, India, Malawi, Nepal, Nigeria, Sudan and Zambia

[73] Of the 65 post-2010 projects operational in those countries listed on the website, 34 (52%) had at least some documents available to be reviewed.

A total of 82 documents were appraised, with the following breakdown.

Document type	Post-2010 programme	Pre-2011 programme	Total reviewed
Business Case	31	5	**36**
LogFrame	25	9	**34**
Annual Review	4		**4**
Planned Evaluation TOR	6		**6**
Previous evaluation		2	**2**
Total			**82**

These documents were read and assessed with a view to looking for ***policy relevance*** (see above), potential ***evaluation questions***, and ***potential project attributes***. The results were collated in an Excel database.

Programme attributes were looked for by way of the following:

5. Outcome (or purpose) and impact indicators (from LogFrames)
6. Descriptions of project outputs, and the weightings of those outputs (from LogFrames)
7. Assumptions in the Assumptions column (from LogFrames)
8. Descriptions of theory of change and proposed intervention (from Business Cases)

Another attribute that could be useful to look for would be the achievement ratings for outputs and outcomes found in Annual Reviews (ARs) and Project Completion Reviews (PCRs) These were not looked for by the evaluability assessment team, however, as only 4 ARs (and no PCRs) were available for the projects looked at.

Limitations to the methodology:

- The DFID website was not always complete. The key statistics page of the DFID website provides overall breakdown of DFID spend into multilateral, bilateral and administration costs, but not specific details within these broad categories.
- Similarly, while the commitment is for 98% of project documents to be online, currently only 40% of projects implemented post-2010 have Business Cases or LogFrames available online (the percentages of documents available for earlier projects are far lower).
- DFID staff only had time to provide policy relevance scoring for a small number of the overall sample of projects looked at by the evaluability assessment team.
- The sample of project projects that were found to be policy relevant was small (15), so percentages drawn from this should be treated with caution.

8.3. Annex C: List of Stakeholders consulted

	Name	Role/Responsibility
	Cindy Berman	Asia Regional SDA – gender focal point for Asia Division
	Jane Doogan	Deputy Team Leader Gender Team, Policy Division, DFID
	Jane Hobson	Reproductive Health Pillar Lead Policy Division, DFID
	Kate Bishop **Kathryn Lockett**	Violence Against Women Pillar Lead, DFID CHASE
	Kate Greany	Gender Team, Policy Division, DFID
	Lindi Hlanze	Economic Asset Pillar Lead Policy Division, DFID
	Rebecca Trafford-Roberts	Adviser on Gender Policy and Evidence Gender Team, Policy Division, DFID
	Ros Ebdon	Team Leader Gender Team, Policy Division, DFID
	Sally Gear	Education Pillar Lead Policy Division, DFID
	Sue Bassett	Regional Policy Adviser for Africa, DFID
	Teresa Durand	Adviser on Gender Policy and Evidence Gender Team, Policy Division, DFID
	Zoe Stephenson	Evaluation Adviser, Gender Evidence and Evaluation Department, DFID
	Name	**Role/Responsibility**
	Daniel Jones	Co-chair of PPA E&A learning group, Christian Aid
	Helen Richards	Governance adviser Governance, Conflict and Social Development Research Research and Evidence Division (RED), DFID
	Isabelle Cardinal	Social Development Adviser Empowerment and Accountability team, Policy Division, DFID
	Jake Allen	Chair of PPA learning group into Measuring Results in E&A, Christian Aid
	Julia Chambers	Civil Society Department, DFID Responsible for PPA and GPAF
	Lina Payne	Evaluation Adviser, Governance & Social Development Evidence & Evaluation Department, DFID
	Lu Ecclestone	Governance adviser Empowerment and Accountability team, Policy Division, DFID
	Shiona Ruhemann	Deputy Head, Governance, Open Societies and Anti-Corruption Dept Policy Division, DFID

8.4. Annex D: Evaluability assessment bibliography

Brown, K. and Van Voorhis, P. (unknown date).*Evaluability Assessment: A Tool for Program Development in Corrections*, for National Institute of Corrections.

Dawkins, N. (2005), *Evaluability Assessments - Achieving Better Evaluations*. PhD, MPH. ICF Macro.

Dun, E. (2008), *Planning for Cost Effective Evaluation with Evaluability Assessment*. Impact Assessment Primer Series, Publication No. 6, USAID.

Lawson, A., Booth, D., Harding, A., Hoole, D., and Naschold, F. (2000) *Evaluability Study Phase 1 Synthesis Report Volume I,* ODI Sida Studies in Evaluation, 00/3

Leviton, L.C., Kettel Khan, L., Rog, D., Dawkins, N., and Coton, D. (2010), Evaluability Assessment to Improve Public Health Policies, Programs, and Practices, *Annual Review of Public Health* 2010.31:213-233. Downloaded from www.annualreviews.org

Leviton, L.C. (2006), *Evaluability Assessment - Practice and Potential*, PowerPoint Presentation, June 14, 2006. Downloaded from www.docsfiles.com/pdf/1/evaluability-assessment.html

Evaluability Assessment, in Evaluating Socio Economic Development, SOURCEBOOK 2: Methods & Techniques (2009)

http://ec.europa.eu/regional_policy/sources/docgener/evaluation/evalsed/sourcebooks/method_techniques/structuring_evaluations/evaluability/index_en.htm

Monitoring and Evaluability Study for the GAVI Alliance Support for CSOs. Final Report and Proposed Monitoring & Evaluation Plan (2008).Prepared by JSI Evaluation Planning Team

Ogilvie, D., Cummins, S., Pettrigrew, M., White, M., Jones, A. and Wheeler, K. (2011), *Assessing the Evaluability of Complex Public Health Interventions: Five Questions for Researchers, Funders, and Policymakers.* Milbank Quarterly, 89: 206–225. doi: 10.1111/j.1468-0009.2011.00626.x

Poate, D., Riddell, R., Chapman, N., Curran, T. et al. (2000), *The Evaluability of Democracy and Human Rights Projects, a LogFrame related assessment,* ITAD Ltd. in association with ODI for SIDA Studies in Evaluation 00/3

Plowman, B., and Lucas, H. (2011), *Evaluability Study of Partnership Initiatives, Norwegian Support to Achieve Millennium Development Goals 4 & 5,* (2010), Mott Macdonald and HLSP, for NORAD Evaluation Department, Report 9/2011

Shadish W.R., Cook T.D., Leviton L.C. (1991), *Foundations of Program Evaluation: Theorists and their Theories.* Newbury Park, CA: Sage

Smith, N.L. (1981), Evaluability Assessment: A Retrospective Illustration and Review, *Educational Evaluation and Policy Analysis* Vol. 3, No. 1, pp. 77-82, Published by: American Educational Research Association Article Stable URL: http://www.jstor.org/stable/1163645

Snodgrass, D., Magill, J., and Chartock, A., Evaluability *Assessment of the USAID/Brazil Micro and Small Enterprise Trade-Led Growth Program* (2006), USAID

Stirling A., (2007), A general framework for analysing diversity, in science, technology and society, J for R. Soc. Interface, 4: 707–719.

Thurston W.E., and Potvin, L., (2003), Evaluability Assessment: A Tool for Incorporating Evaluation in Social Change Programmes, *Evaluation*, 9:4, pp 453–469; SAGE Publications, London.

Trevisan, M. S. and Yi Min Huang (2003), Evaluability assessment: a primer, Practical *Assessment, Research & Evaluation*, 8(20)

UNIFEM (2010) Terms of Reference for an Evaluability Assessment of the UNIFEM Strategic Plan 2008-2013.

Other documents referenced in report

Final Report-Learning from DFID's Governance and Transparency Fund (GTF): Tools, methods and approaches
Richard Burge, 4th June 2010. DFID. TripleLine, KPMG.

How to Sort, by Harloff and Coxon, 2009 and their associated The Method of Sorting website. For online sorting, see WebSort.net and OptimalSort

Scriven M. The Final Synthesis. Sage, American Journal of Evaluation, 15/3(1994):367-82.

Davies, R, (2011) 3ie and the Funding of Impact Evaluations. A DISCUSSION PAPER FOR AUSAID, page 8.
Data Mining with Decision Trees: Theory and Applications by Lior Rokach, Oded Z. Maimon. World Scientific, 1 Mar 2008

Modelling. Christina H. Gladwin. Sage, 1989

Funnel, S., Rogers, P. (2011) Purposeful Program Theory: Effective Use of Theories of Change and Logic Models. Jossey-Bass

McGee and Gaventa (2011) Shifting Power? Assessing the implications of Transparency and Accountability Initiatives, IDS Working Paper Volume 2011 No. 383

Rondinelli, D.A. (1993) Development Projects as Policy Experiments, Routledge

Shmueli, G; Patel, N.R., and Bruce, P. C. Bruce (2010).Wiley Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner

http://mandenews.blogspot.co.uk/2012/04/criteria-for-assessing-evaluablity-of.html

http://mande.co.uk/2011/lists/value-for-money-a-beginners-list/ [Editor's Suggestions] and http://www.mandenews.blogspot.co.uk/2012/05/perspective-on-value-for-money.html

http://www.mandenews.blogspot.co.uk/2012/06/open-source-evaluation-way-forward.html

http://mandenews.blogspot.co.uk/2010/10/do-we-need-minimal-level-of-failure-mlf.html

http://www.3ieimpact.org/userfiles/doc/SR_blog.pdf, accessed in April 2012

8.5. Annex E: Proposed process for the card sorting exercise with DFID

Background references to card sorting:

- Hierarchical Card Sorting, Rick Davies,1996 and later
- How to sort: A short guide to sorting investigations, *Harloff and Coxon, 2007*

Objectives

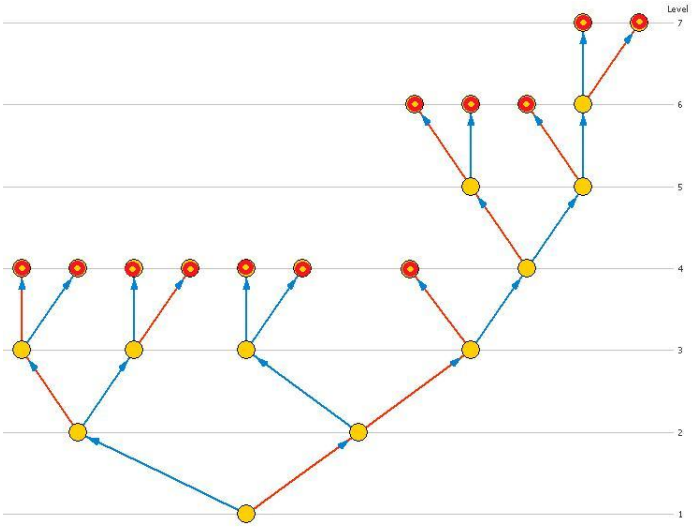
1. For each policy area, identify a **sample** of up to 7 countries out of a total set of 28 in which DFID is working, which can be the focus of the evaluability assessment, and
    - To do so in a way that maximises their diversity (explained further below)
2. Generate hypotheses, which can be phrased as evaluation **questions**, about "what works where" that might prove to be evaluable and of interest to macro-evaluation stakeholders
    - These may also inform revisions to the generic ToC for the policy area (E&E/W&G)

Proposed process

1. Identify the items to be sorted. These are the 28 focus countries where DFID is working , as listed on its website, where there may be programs focusing on E&A and/or W&G
    - *Afghanistan, Bangladesh, Burma, DR Congo, Ethiopia, Ghana, India, Kenya, Kyrgyzstan, Liberia, Mozambique, Malawi, Nigeria, Nepal, Occupied Palestinian Territories, Pakistan, Rwanda, Sierra Leone, South Africa, South Sudan, Sudan, Somalia, Tanzania, Tajikistan, Uganda, Yemen, Zambia, Zimbabwe.*
        - This list needs checking by DFID
        - Then their names written on prominently on file cards, one country per card
2. Establish the sorting instructions. This will be in the form of a single question as follows:
    - "*Please sort these countries into two piles, of any size, according to what you think is the most significant difference between them. We are interested in differences that you think make a difference, to how the E&A/W&G programs in those countries are working*"
3. Document the results
    - When the cards have been sorted into two piles, document which cards/countries are in which pile, and ask for an explanation of the difference and the difference it makes to the implementation of the program. Feed the explanation back to check it has been understood
4. Re-iterate the process
    - Taking one of the first two piles, ask the same sorting question again, and again document the results. Do the same with the second pile,
    - Repeat the process until all piles that have been generated consist of only one card. Or stop when the respondent says they can't identify a significant difference.

Comments on the process and results

1. The "differences that make a difference" are in effect hypotheses or expectations, which can be phrased as questions that an evaluation could ask
    - If there is time left over, it would be useful to ask respondents which of the hypotheses would be most useful to try to test via one or other evaluations. Bearing in mind that:
        - Some may already be proven, in the sense that evidence can already provided
        - Some may not be provable, because it would difficult to find appropriate evidence
2. From amongst the 28 countries we need to pick a sample of up to 7 (as discussed on 28 March) that are maximally different. The results of the card sorting can be visualised as a tree structure. Here is one from a previous exercise.



The red circles are the sorted cards, and the yellow circles are the groups they were placed in from the beginning of the sorting exercise, starting at the bottom

The proposed process for selecting a sample (say of 6 countries out of the above 12) is very simple. It is to select every 2nd country from left to right. This will mean that types of countries will be represented in proportion to their numbers of members. In the above example, the right side main branch will have four countries and the left side will have two.

- If we needed to quantify the diversity contained within the sample we can calculate the average distance between each of the sample countries, where each link in the chain of links connecting them is a unit of measurement (known as degree, in social network analysis). This average is an aggregate measure known as "closeness". The results of other sample choices could be compared using this simple metric[74].

In discussions with DFID it has been proposed that the overall sample of countries will also need to be checked to make sure that in aggregate it is not *unintentionally* skewed e.g. is the average GNP of the sample countries higher than the whole set of 28 countries.

[74] For those interested, this corresponds to Stirling's concept of "disparity" in his analysis of the measurement of diversity.

Caveat

It was originally proposed that the sorting exercise would focus on **comparisons of E&A/W&G programs** in the 28 countries. On reflection this may not be advisable, for two reasons. 1. This approach demands more detailed knowledge than a **comparison of the countries** themselves, 2. The programs in each country may overlap in kind, making comparisons more difficult than comparisons of discrete countries. On the more positive side, comparisons of **countries** will help generate hypotheses about the *contexts* in which E&A/W&G programs are operating. Context is an important part of a good ToC, but often neglected, relative to the attention given to interventions. This view is argued by the "realist evaluation" school, which is well known for its useful phrase: *context + mechanism = outcome*

Hypotheses about the importance of different mechanisms could be identified later on by enquiries within each of the sampled countries. If an appropriate respondent could be found they could be given a variant of the above sorting instruction e.g.:

> *"Please sort these programs into two piles, of any size, according to what you think is the most significant difference between them. We are interested in differences that you think make a difference, to what these programs are able to achieve in this country"*

The answers could also inform sampling choices by a macro-evaluation, if there are too many programs to examine as a whole.

8.6. Annex F: Results of country card sorting exercises

Note

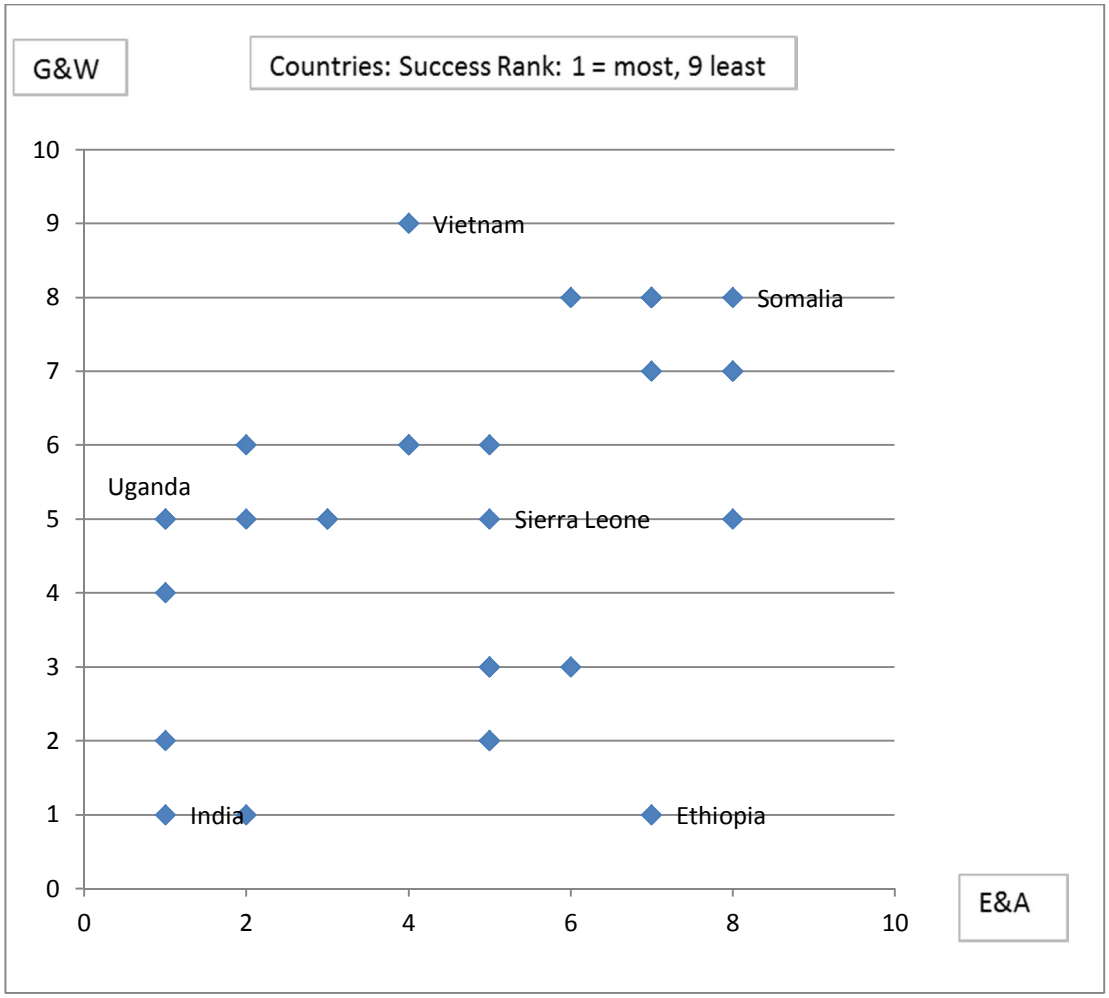
1. The tables below describe the most significant differences identified between the groups of countries, but not the differences they made (the expected or observed consequences)
2. The branches have been ordered into their implied ranking, by making an assumption at each branch point as to which group would produce better results. This ranking could be contested and changed by the participants. The current or revised ordering could be treated as testable claims about expected relative success in different contexts
3. Yellow countries are outside the DFID 28 focus group countries. Vietnam was not included in the E&A card sort exercise

VW&G policy team results

ID W&G	Differences that make a difference				Country	Success ranking
1	More committment from senior DFID staff in country	More innovative	More conducive enabling envionment	Bigger programmes	India	1
2	More committment from senior DFID staff in country	More innovative	More conducive enabling envionment	Bigger programmes	Bangladesh	1
3	More committment from senior DFID staff in country	More innovative	More conducive enabling envionment	Bigger programmes	Ethiopia	1
4	More committment from senior DFID staff in country	More innovative	More conducive enabling envionment	Smaller programmes	Ghana	2
5	More committment from senior DFID staff in country	More innovative	More conducive enabling envionment	Smaller programmes	Nepal	2
6	More committment from senior DFID staff in country	More innovative	More conducive enabling envionment	Smaller programmes	Rwanda	2
7	More committment from senior DFID staff in country	More innovative	Less conducive enabling environment	Bigger programmes	Pakistan	3
8	More committment from senior DFID staff in country	More innovative	Less conducive enabling environment	Bigger programmes	DRC	3
9	More committment from senior DFID staff in country	More innovative	Less conducive enabling environment	Bigger programmes	Nigeria	3
10	More committment from senior DFID staff in country	More innovative	Less conducive enabling environment	Smaller programmes	Zambia	4
11	More committment from senior DFID staff in country	Less innovative			Tanzania	5
12	More committment from senior DFID staff in country	Less innovative			Uganda	5
13	More committment from senior DFID staff in country	Less innovative			Kenya	5
14	More committment from senior DFID staff in country	Less innovative			Mozambique	5
15	More committment from senior DFID staff in country	Less innovative			Malawi	5
16	More committment from senior DFID staff in country	Less innovative			Sierra Leone	5
17	More committment from senior DFID staff in country	Less innovative			Zimbabwe	5
18	Less commitment by senior DFID staff in country to G&W Vision		More stable states		South Africa	6
19	Less commitment by senior DFID staff in country to G&W Vision		More stable states		Kyrgystan	6
20	Less commitment by senior DFID staff in country to G&W Vision		More stable states		Tajikistyan	6
21	Less commitment by senior DFID staff in country to G&W Vision		More stable states		Burundi	6
27	Less commitment by senior DFID staff in country to G&W Vision		Fragile states	More access	Sudan	7
28	Less commitment by senior DFID staff in country to G&W Vision		Fragile states	More access	Burma	7
29	Less commitment by senior DFID staff in country to G&W Vision		Fragile states	More access	Sudan	7
22	Less commitment by senior DFID staff in country to G&W Vision		Fragile states	Poor access by DFID staff & consultants	Liberia	8
23	Less commitment by senior DFID staff in country to G&W Vision		Fragile states	Poor access by DFID staff & consultants	Yemen	8
24	Less commitment by senior DFID staff in country to G&W Vision		Fragile states	Poor access by DFID staff & consultants	Afghanistan	8
25	Less commitment by senior DFID staff in country to G&W Vision		Fragile states	Poor access by DFID staff & consultants	OPT	8
26	Less commitment by senior DFID staff in country to G&W Vision		Fragile states	Poor access by DFID staff & consultants	Somalia	8
30					Vietnam	9

E&A policy team results

ID E&A	Differences that make a difference			Country	Success ranking
1	Less fragile	NSAs have more capacity	budget support	Zambia	1
2	Less fragile	NSAs have more capacity	budget support	Ghana	1
3	Less fragile	NSAs have more capacity	budget support	Tanzania	1
4	Less fragile	NSAs have more capacity	budget support	Uganda	1
5	Less fragile	NSAs have more capacity	budget support	India	1
6	Less fragile	NSAs have more capacity	non budget support	Kenya	2
7	Less fragile	NSAs have more capacity	non budget support	South Africa	2
8	Less fragile	NSAs have more capacity	non budget support	Bangladesh	2
9	Less fragile	less capacity NSAs	DFID higher priority	Mozambique	3
10	Less fragile	less capacity NSAs	DFID higher priority	Malawi	3
11	Less fragile	less capacity NSAs	DFID Lower priority	Vietnam	4
12	Less fragile	less capacity NSAs	DFID Lower priority	Kyrgystan	4
13	Less fragile	less capacity NSAs	DFID Lower priority	Tajikistan	4
14	Fragile	More space for NSAs	budget support	Burundi	5
15	Fragile	More space for NSAs	budget support	Nepal	5
16	Fragile	More space for NSAs	budget support	Pakistan	5
17	Fragile	More space for NSAs	budget support	DRC	5
18	Fragile	More space for NSAs	budget support	Sierra Leone	5
19	Fragile	More space for NSAs	budget support	Rwanda	5
20	Fragile	More space for NSAs	No budget support	Nigeria	6
21	Fragile	More space for NSAs	No budget support	Liberia	6
22	Fragile	Less space for NSAs	More Engagement	South Sudan	7
23	Fragile	Less space for NSAs	More Engagement	Yemen	7
24	Fragile	Less space for NSAs	More Engagement	Afghanistan	7
25	Fragile	Less space for NSAs	More Engagement	OPT	7
26	Fragile	Less space for NSAs	More Engagement	Ethiopia	7
27	Fragile	Less space for NSAs	Little Engagement	Burma	8
28	Fragile	Less space for NSAs	Little Engagement	Sudan	8
29	Fragile	Less space for NSAs	Little Engagement	Somalia	8
30	Fragile	Less space for NSAs	Little Engagement	Zimbabwe	8



Of note:

1. India expected to be most successful in terms of E&A and SVG&W
2. Somalia expected to be least successful in terms of E&A and SVG&W
3. Ethiopia expected to be most successful in terms of G&W but least successful in terms of E&A
4. Uganda expected to be least successful in terms of G&W but most successful in terms of E&A

8.7. Annex G: Attributes of planned evaluations with policy relevance

Note: This table is based on analysis of the scheduled evaluations listed in the 'stocktake of planned evaluations' provided by DFID's EvD to the evaluability assessment team in April 2012. Of the 340 scheduled evaluations listed in the stocktake, 116 (34%) of the evaluations were identified as policy relevant (74 E&A, 42 SVG&W) by DFID's Evaluation Department.

The table below shows a breakdown of these evaluations by type and management arrangements.

Type of evaluation	**E&A**	**SVG&W**
Impact evaluations	28%	48%
Process evaluations	17%	10%
Policy evaluations	7%	
Thematic evaluations	5%	2%
Other	20%	14%
Not yet clear	21%	26%
Management		
DFID only	35%	33%
Joint	33%	21%
Not yet clear	32%	46%
Stage		
Summative	33%	31%
Formative	26%	21%
Not yet known	41%	48%
N	74	42

8.8. Annex H: Analysing categorical data and visualising the results

Aspects of the context, interventions and outcomes can all be described using binary categories, and whole projects can be described using multiple sets of categories. As with the use of more sophisticated measures, care must be taken to carefully apply such categories[75].

There are at least four methods of analysis that can use categorical data, to identify relationships between events:

1. Qualitative Comparative Analysis. This is a theory-led hypothesis testing approach, suitable to small numbers of cases. It is now becoming better known in some evaluation circles.
2. Data mining algorithms used for the discovery of association rules (represented as Decision Trees or Classification Trees). This is an inductive approach, complementary to hypothesis testing, and can be used on small and large numbers of cases. This is widely used in business and biological science, but relatively unknown and unused by evaluators[76].
3. Ethnographic Decision Tree Modelling, a participatory approach usually developed with small numbers of cases but testable on large numbers. This approach has been around since the 1980s, but remains a niche interest[77].
4. Hierarchical Card Sorting, another participatory approach which produces decision trees that can be used for multiple purposes[78]. Card sorting methods are a well-known ethnographic tool, but they are not widely used to generate decision trees

All four methods can generate testable explanations of cases, and testable predictions of results that will be found when applied to sets of new cases (having the same categories of attributes). Non-parametric statistical tests can be applied to the results, the most well-known of which is probably the Chi-Square test.

These methods have four other advantages

- They can identify and describe multiple configurations of attributes[79]that are associated with specific outcomes, involving attributes that may or may not be necessary and/or sufficient causes. Examples can be seen online[80].
- The methods assume and exploit heterogeneity of interventions and contexts, in contrast to experimental methods which assume homogeneity (or limited heterogeneity at best)
- Categorical data used by these methods can be generated by both participatory and expert means.
- The results of all four methods can be graphically represented as Decision Tree diagrams, which are easy to read and understand. The raw data can be held in simple Excel files.

[75] To ensure construct validity (you are describing what you think you are describing) and measurement reliability(others use the category in the same way you do)
[76] Data Mining with Decision Trees: Theory and Applications by Lior Rokach, Oded Z. Maimon. World Scientific, 1 Mar 2008
[77] Modelling. Christina H. Gladwin. Sage, 1989
[78] See http://mande.co.uk/special-issues/hierarchical-card-sorting-hcs/
[79] This capacity is needed to address the problem of "equifinality" – the possibility that there may be multiple paths or combinations of variables/attributes that can produce the same kind of outcome. Also described as the problem of over determination - An event is over determined if there exist more than one antecedent events, any of which would be a sufficient condition for the event occurring.
[80] http://mandenews.blogspot.co.uk/2012/06/representing-different-combinations-of.html

At the level of individual projects it is also likely that there will be interval and ratio scale data that is amenable to conventional statistical analyses. For example, via baseline and follow-up surveys. This information could be used during individual evaluations for more in depth analysis, treating the project concerned as a case study opportunity.

Case studies have an important function in relation to each of the methods described above. Each of the methods can identify *associations* between different project attributes and outcomes, found across a set of projects, many of which may appear to reflect plausible causal connections. However in-depth investigations of individual cases, e.g. via a project specific evaluation, can provide an opportunity to find out if a plausible causal connection is working as expected. They can also be used to check if attributes have been correctly ascribed to a given project.

This approach is consistent with arguments made by others that claims of causal attribution need to be made by combining two kinds of evidence[81]:

- Co-variance data: showing how an apparent cause and event co-occur and are mutually not present.
- Mechanism explanations: showing how the cause is expected to lead to the effect.

[81] Points made here are well summarised in two sources: B. Befani (2012) *Models of Causality and Causal Inference*, (especially the final section) and *Good Thinking* by Denise Cummins, 2012, CUP, chapter 6 What causes what"

8.9. Annex I – Comment on Evaluation Questions

Annex I: Comments on evaluation questions

Background

This Annex should be read in conjunction with section 5 of the main Evaluability Assessment report

The Annex is in two parts, developed at different times and as a result they have different structures

Part 1: Evaluation questions concerning DFID's Strategic Vision for Girls and Women

Text in normal font is taken from Annex 1 to Evaluability Assessment Terms of Reference
Text in indented italic font are comments by the Evaluability Assessment team

Comments have been made on the SVG&W questions only for two reasons: (a) They are more numerous and detailed, (b) the over-arching questions asked by the E&A team are very similar in type to the categories of questions discussed in section 5 of the Evaluability Assessment report

i) Questions on Overall Strategy

- Were the four pillars the most strategic and effective entry points through which to promote girls and women' empowerment? *(DAC criteria: relevance)*
    - *Methodological assessment*
        - *Does "entry point" = approach?*
        - *If so, what other approaches would be the comparator?*
            - *Option 1: Rate projects according to their fidelity to the G&W Vision, or*
            - *Option 2: Assume newest projects have higher fidelity than older projects*
        - *And how is "most effective" to be assessed?*
            - *PCR ratings? There will be no other common measure*
    - *Practical assessment*
        - *If the answer was not so, what would happen? Would the 4 pillars approach be abandoned? This seems unlikely*
        - *Or, could the SVG&W ToC be adapted to incorporate functional approaches currently outside the Vision. More likely*
    - *Possible action: Incorporate fidelity ratings in the proposed projects database and extend analysis of PCR ratings proposed in section 5*
- What impact has the Strategic Vision had on DFID's relationship with external partners?*(relevance)*
    - *Methodological assessment*
        - *External partners could be surveyed, as to their knowledge of and attitude to the SVG&W*
    - *Practical assessment*
        - *If they knew nothing or did not like it, would this lead to:*
            - *A change in the DFID SVG&W – least likely*
            - *Some new communication initiative - possible.*
            - *Or not change at all – possible*
    - *Possible action: Treat as a low priority*

- To what extent was the Strategic Vision design coherent, logical and innovative (*Effectiveness)*
    - *Methodological assessment*
        - *Stakeholder views within and outside DFID would need to be surveyed. There is no objective basis for judgement*
        - *Sampling would be critical, otherwise results could be seen as inevitably biased (even if not so in practice)*
    - *Practical assessment*
        - *Results could affect how the Vision is subsequently communicated*
    - *Possible action: Treat as a low priority or incorporate within proposed policy implementation review*

- What effect did the Vision's particular focus on girls have on DFID programmes? (*Effectiveness)*
    - *Methodological assessment*
        - *This would require a comparison of projects pre and post SVG&W*
            - *This could be a desk study*
            - *Doing so would require a sample of DFID projects, possibly by most relevant OECD/DAC input codes*
        - *More specific and testable views on possible effects are needed*
            - *Might best be done via staff surveys*
    - *Practical assessment*
        - *A "significant change" finding would be very acceptable, but would not imply need for any change.*
        - *A "no significant change" finding could be embarrassing. Would that lead to any change – uncertain. So VfM of this question may be in doubt*
    - *Possible action: Treat as a low priorityor incorporate within proposed policy implementation review*
- Has Value for Money been achieved in delivering the vision?
    - *Methodological assessment*
        - *Comparison would best be with pre-Vision projects*
            - *2nd best would be low fidelity post 2010 projects*
        - *Comparisons of their value would only be possible for clusters of projects with comparable outcomes*
        - *Only gross costs data is likely to be available for pre-Vision projects now completed*
            - *Pre-and post average total costs could be compared*
            - *Some form of unit costing may be possible*
    - *Practical assessment*
        - *It could be more useful to analyse VfM differences within the current portfolio of Vision projects, where access to cost data would be better, and results could inform new project designs*
    - *Possible action: Treat as a low priority, in its current form*

ii) Questions on Results

- To what extent are the results achieved by the vision quantifiable and measurable?
    - *This question is answerable*
    - *Possible action: See EA report section 4 on Theory of Change and section 3 on data availability*
- Did the Vision achieve the results it set out to? (*Impact? Effectiveness?)*
    - *Answers will be available in cumulative "We Wills" data, but this is only a partial picture of what the Vision seeks*

        - *The two indicators at the top of the House model are not being comprehensively monitored across all policy relevant projects*
    - *Possible action: PCR ratings may be usable proxy measures of overall achievement, and they will be available. See section 5 on Evaluation Questions*
- What impacts has the Vision had on empowerment of girls and women? (*Impact*)
    - *Answers will require a consistent definition of empowerment*
        - *This could be single or composite (weighted) measure*
    - *And new measurement of all projects in these terms, where a consistent measure is not in place already*
    - *Possible action: In their absence project PCR ratings might be used as a proxy of empowerment. See section 5 on Evaluation Questions*
- Was value for money achieved in delivering these results, at all stages of the results chain?
    - *This is not possible to answer. It would require comprehensive access to costs information from a wide range of partnersover an extended period of time*
    - *Possible action: Do not try to answer this question*
- Do results depend on interaction between pillars? What difference does it make to girls and women if they benefit from more than one area of the Vision? *(Impact and Effectiveness?)*
    - *This question should be answerable.*
    - *Possible action: See section 5 of main report*
- Do impacts depend on progress in the enabling environment? (*Impact)*
    - *Interpreted literally, the answer would almost inevitably be yes. No need for an evaluation*
    - *The question would be more useful and evaluable if the concept of "enabling environment" was disaggregated into some observable elements.*
    - *Possible action:Evaluation of the importance of different context features is possible and is discussed in section 5*
- To what extent are the results achieved to date and future results likely to endure into the longer term? *(Sustainability)*
    - *This seeks a global generalisation, which will be difficult to make at the best of times*
        - *It would be better to seek differences in likely sustainability*
    - *These would then need to be tested by follow up work, say 1 year after project completion. This would fall into the category of commissioned evaluations discussed in section 6.*
    - *But it is doubtful if commitments made to do so now will be honoured by others some years down the line.*
    - *Possible action: A low priority*

iii) Questions Implementation of the Vision

- How has the Vision guided the work of DFID Country Offices on girls and women? (*Effectiveness)*
    - *This is evaluable. Staff could be surveyed, and their claims verified by analysis of project documentation.*
    - *The results should be practically useful*
    - *Possible action: Via a Policy Implementation Review. See section 5 of report.*
- How has implementation varied across the organisation? *(Effectiveness and Efficiency)*
    - *This is evaluable. This could be done by analysis of project documents in the first instance, then interviews of staff*
    - *The results should be practically useful*
    - *Possible action: Via a Policy Implementation Review. See section 5 of report*

- Are there effective cross-pillar linkages? Are there effective linkages between the four pillars and the enabling environment? If so how have these been achieved? *(Effectiveness and Efficiency)*
    - *Potentially evaluable, if views on likely "effective cross-pillar linkages" can be identified and treated as testable hypotheses*
    - *Results could make a difference. Absence of linkages could be cheaper and easier to implement, but may have less effect.*
    - *Possible action: Via database analysis and scheduled evaluations. See section 5 of report*
- Have different approaches to implementation affected the extent to which value for money has been achieved*?*
    - *Interpreted literally, the answer would almost inevitably be yes. No need for an evaluation*
    - *Possible action: Section 5 proposes a crude analysis involving comparison of project costs with their PCR ratings.*
        - *The most useful result will not be the correlation coefficient, but case studies of the outliers (i.e. high cost failures and low cost successes)*

iv) Questions on the institutional arrangements

- Do the organisational structures for the Strategic Vision provide clear leadership, a strong accountability structure and positive incentives for effective delivery of DFID's work on girls and women? *(Effectiveness and Efficiency)*
    - *This is very much about policy implementation within DFID.*
    - *A conventional project centred evaluation would not be applicable*
    - *Possible action: An externally facilitated internal review would be most appropriate*
- Has the Vision led to an increase in the allocation of financial resources to programmes on girls and women?
    - *This is evaluable*
    - *Possible action: An analysis of data from ARIES should be sufficient*
    - And increased mainstreaming of girls and women in DFID programmes?
        - *This would be evaluable if mainstreaming could be defined and measured, possibly by the use of a project rating instrument*
        - *It would then retire comparison of new and old projects, which could be done using DFID databases*
    - If so, has our ability to track spending on girls and women changed as a result of the vision? *(Effectiveness)*
        - *This may be evaluable, but requires knowledge of the structure of DFID databases*
        - *The findings would be useful, hopefully leading to changes in database structures and use*
        - *Possible action: Not known*
- How have reporting requirements in the Corporate Performance Framework affected implementation of the Vision? *(Effectiveness)*
    - *This is potentially evaluable*
    - *Possible action: Answers could be sought through staff surveys as proposed above, as part of a policy implementation review*
        - *And claims checked against project documentation*
- How effectively did DFID respond to risks identified in the Vision and to changes (opportunities and challenges) in the external environment? (*Effectiveness)*

- This is potentially evaluable
- The main instrument would be staff interviews
- The challenge will be in verification of recollections of risks identified and responses made

Part 2: Evaluation questions concerning the Empowerment and Evaluation policy area

An early draft of proposals for evaluable evaluation questions. Based on EA TOR list of types of "over-arching evaluation questions" Underneath these are supposed to come:

- Subsidiary questions that can be utilised by country programme
- A purely local set of questions not related to the meta-evaluation

1. *The overall impact of increased work on E&A: e.g. have efforts on E&A made any difference to poverty, development, fragility, governance outcomes, or to social cohesion or power relations?*
    - Candidate Evaluable Claim 1: **The majority of DFID projects which are "largely focusing on" the E&A policy objectives have achieved their objectives by the end of their planned term.**
        - "Largely focused on" or "wholly focused on" as described by the proposed policy relevance rating scale. Such a rating would need checking by a MA team
        - "E&A policy objectives" as expressed in the "Strengthening Empowerment and Accountability in International Development: Emerging Guidance" paper. There needs to be a common reference point for such judgements
        - Achievement of objectives as described by a rating of A, A+ or A++ provided in Project Completion Review, unless contradicted by an independent evaluation. This is the only common measure of achievement across all kinds of E&A (or other) projects.
            - Evaluations could be both a source of validation for the PCR judgements and a means of doing more in-depth inquiry into the causal processes involved. Including examination of outliers, the contradictory cases, where lessons may be learned
                - PS: Often PCRs are completed as part of an independent evaluation

2. *Impact of specific interventions in different contexts (what works and what doesn't and why?) Why have some programmes worked well for some groups but not others?*
    - Candidate Evaluable Claim 2: **E&Aprojects have achieved more in the following country contexts**
        - **In less fragile states, rather than more fragile states**
        - **Where there is more space for non-state actors rather than less**
        - **Where non-state actors have more capacity rather than less**
        - **Where DFID is providing budget support, versus those where it is not**
        - **Where all these conditions (the former) are combined rather than where all these conditions are absent (the latter)**
    - The countries that fall into these categories have been defined by the card sort exercise. These could be revised prior to the MA, as could the set of distinctions used to sort them into groups
    - Achievement as described above

- Candidate Evaluable Claim 3**:** (Same kind of claim as above, but specific to a sampled country with many E&A focused projects, differentiated by a card sort or other means, used to generate a similar set of claims about contextual differences)

3. *Interaction between different programmes working on E&A. For example, does working on a number of areas within the policy frame lead to better results than working on individual programmes – what is the sum of the parts? Does empowerment in one sphere (e.g. economic) lead to accountability in others (e.g. political) or vice versa?) Do interventions that combine both empowerment and accountability elements achieve better development results?*

- Candidate Evaluable Claim 4**:** **Projects that combine both empowerment and accountability elements achieve better development resultsthan those which don't**
-
    - These projects would be those tagged as doing so, by a document search prior to a PCR and or evaluation of the same projects
    - Two classes of these projects could be tagged: (a) where the project itself is providing both elements, (b) where the project plans to cooperate with others who are proving the second element
    - Achievement could be defined as above
- Candidate Evaluable Claim 5**:** **Projects that combine both E&A and G&W elements achieve better development results than those which don't**
    - Conditions as above

4. *Institutional arrangements for supporting E&A work: e.g. is it more effective to support 'stand-alone' voice and accountability interventions, or to integrate E&A into other programmes?*

- Candidate Evaluable Claim 6: **Projects which are only partly focused on E&A policy objectives have achieved their objectives at least to the same extent as those largely focused on those objectives.**
    - Same conditions apply as for Claim 1.
    - It would also be necessary to check that the project Outputs relating to E&A were achieved.

5. *Mid-level strategies on E&A: e.g. What should be the priorities in different contexts? Do some strategies work better in certain contexts and can we generalise about these contexts (build typologies)?*

- Candidate Evaluable Claim 7: **There are some countries where E&A is supposed to do well, but not G&W (Uganda) and some where G&W are expected to do well but not E&A (Ethiopia).** (Implications if so: one strategy may need to be pursued through the other, in some countries)
    - According to results of the card sorting exercise (which can always be revisited and revised)

6. *Drawing out implications for the future: for example, what can we do to improve the success of interventions now underway? How can this intervention be scaled-up or diffused to other settings?*

- Candidate Evaluable Claim 8: The examination of contradictory cases, in the inquiries about each of the above claims, will generate lessons relevant to future project design and policy positions. E.g.
    - If "*The majority of DFID projects which are "largely focusing on" the E&A policy objectives have achieved their objectives by the end of their planned term*" then examine the same kind of projects which did not achieve their objectives by the end of their planned term.
    - If "*E&A projects have achieved more in the following country contexts: In less fragile states, rather than more fragile states*" then examine projects in a more fragile state

that have done relatively well. Or projects in a less fragile state that have done relatively poorly

8.10. Annex J: Types of explanations

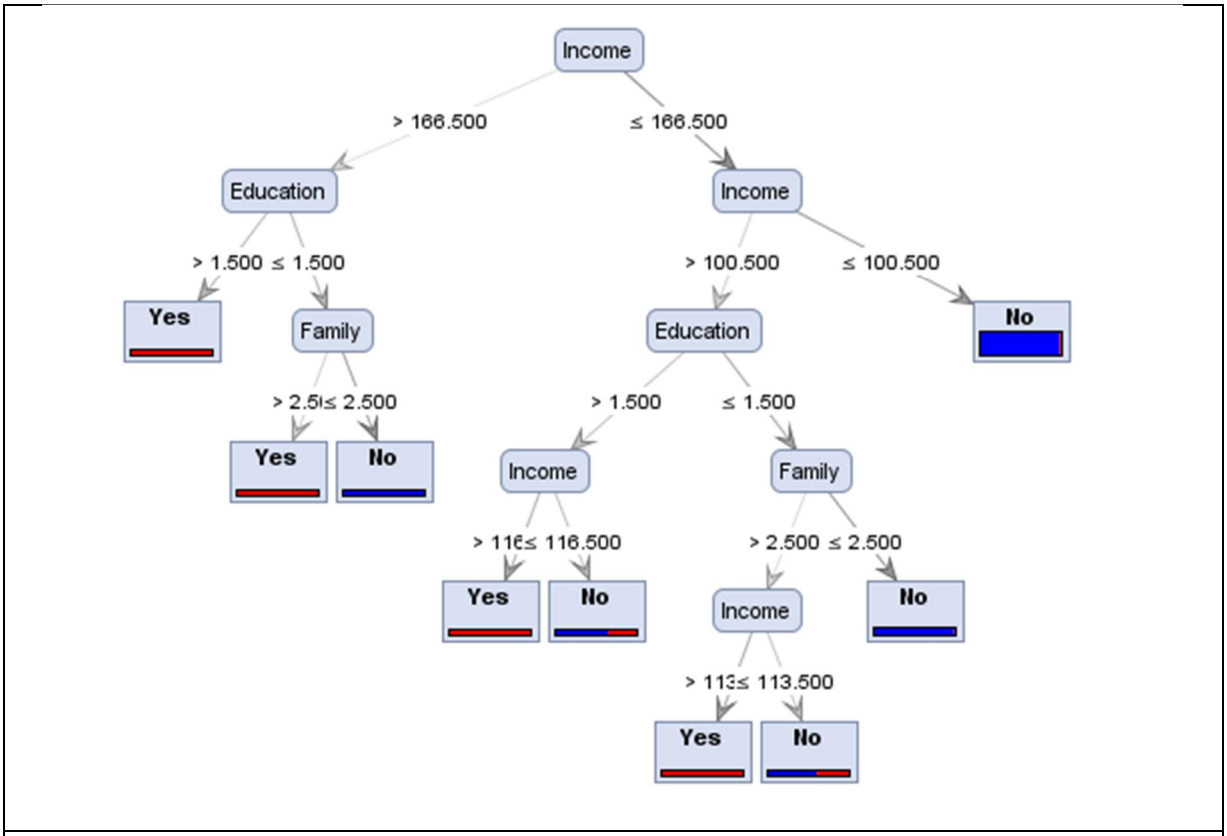
Evaluation processes should be means to an end, not an end in themselves. Ideally answers to evaluation questions will help to accumulate a body of evidence based and testable knowledge about what works in what circumstances SVG&W. To help make sure this happens it is worth considering the types of explanations that answers to evaluation questions could deliver. They could include:

- Single factor explanations: IF X intervention is present THEN Y outcome occurs. For example, the card sorting exercises suggested that there will be more E&A project success in more stable states (versus less stable states), and more SVG&W project success where there is an enabling environment (versus where there is not). While this prediction may be true for some stable states, it is unlikely to be true for everyone. While valuable from a project design point of view, single factor explanations tend to be hard to find for many complex social outcomes.
- Compound explanations: If X context is present and Y intervention takes place THEN Z outcome occurs. This is the classic Realist Evaluation school formula (Context + Mechanism = Outcome). As discussed above, the challenge here will be to develop ways of categorising or tagging different contexts associated with different project interventions. Project documents we have scanned provide relative good descriptions of project activities but poorer descriptions of context.
- Multiple compound explanations. Experiences with the use of Qualitative Comparative Analyses to analyse political and social developments in relatively small number of cases (countries or organisations) suggests that often multiple explanatory rules are needed to explain the full set of observed outcomes. Participatory analyses can generate the same kinds of results. For example the E&A country card sorting exercise result provide 8 different explanations that might account for the outcomes in 28 countries. Some illustrative examples:
    - IF a country is more stable, AND the non-state actors have more capacity than elsewhere, AND E&A initiatives operate within the context of direct budget support THEN there will be greater impact (than in countries where this combination of conditions is not present)
    - IF a country is less stable AND there is less space for non-state actors AND there is little engagement with government THEN there E&A initiatives will have less impact (than in countries where this combination of conditions is not present)

Within compound explanations there may or may not be some *necessary* conditions. For instance, the Policy Division gender team mentioned, during interview as part of this evaluability assessment, that*" A really good gender analysis is necessary for a program to work, of social norms and the enabling environment, the legislation and policy framework, budgets available".* Identification of necessary conditions present in any proposed explanation is important for two reasons. Firstly, it is a point of vulnerability in the explanation that needs testing. Secondly, if found to be true, it needs to be recognised as such and inform the design of subsequent projects.

There may also be other conditions whose presence can be described as sufficient. Their presence has implications for the replicability of interventions. Where an intervention is both necessary and sufficient then strong claims can be made about attribution, about having caused the outcome.

The best way to summarise multiple compound explanations, is in the form of decision trees, of the kind shown in below (a business application). Note that this does not require commitment to a specific evaluation or research method but simply the use of a particular way of summarising the results in a readable and testable form



Decision Tree example: Descriptive model of kinds of people that accept offers of personal loans

Blue = cases of people that reject offers. Red = cases of people that accept

The numbers on each branch refer to the cut-off points used to distinguish different levels of income, education, etc.

There are four groups of acceptors whose behaviour is each explained by a different combination of conditions. The dataset contains 12 customer attributes ranging from Income, Education in years, mortgage, average credit card balance, family size, and geographic data among others. But the explanatory model only needed to use three attributes

Source: http://www.simafore.com/blog/bid/94454/A-simple-explanation-of-how-entropy-fuels-a-decision-tree-model

Decision trees representations have three kinds of merit:

1. They can be constructed by different means, ranging from computerised analysis of large data sets to ethnographic enquiries about the practices of individual people. They can make use of nominal, ordinal, interval or ratio scale data. They can be constructed for any number of cases, small or large.
2. The workings of decision trees are transparent and user friendly, relative to many other ways of summarising knowledge. E.G. Boolean logic expressions in QCA, or the results of Regression analyses

3. They are testable. Decision trees developed as good description of one set of cases can be tested for their predictive accuracy against the same kind of outcomes observed in another set of cases.

Caveat: Explanations given in the form of rules (IF X is present AND Y is done THEN Z happens) are all about observed *associations*. Like correlations between variables they may or may not represent real causal processes. There may be circumstances in development projects where prediction is sufficient and causal attribution is not needed as well. For example, in the design of immunisation campaigns that delivers the highest levels of coverage. But for the design of new projects based on analysis of past projects, confidence about causal processes at work will be important. One way of finding out is to do case studies of individual cases, to see if their specific internal workings are consistent with the way the general rule suggests things are working. Scheduled evaluations could provide opportunities for such case studies.