Macro evaluation of DFID's Policy Frame for Empowerment and Accountability

Methodology

A.1 Introduction

The macro evaluation methodology to understand from DFID's project portfolio what works, for whom and in what contexts requires work at two levels (Figure 1):

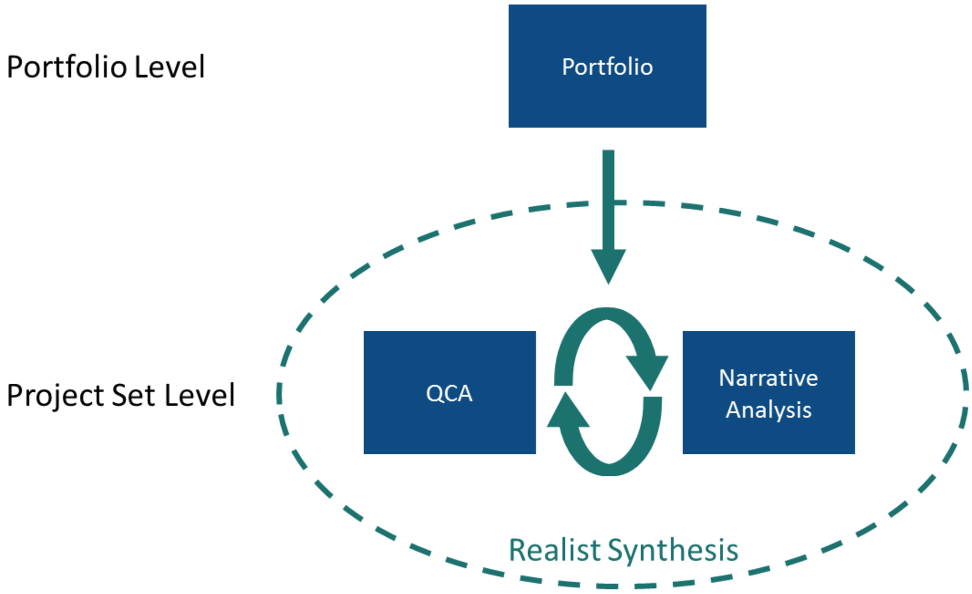
- The Empowerment and Accountability (E&A) project portfolio level;
- The project set level, with a set being a cluster of E&A projects with a common intended outcome.

DFID project management systems do not facilitate the comprehensive identification of all projects relevant to the E&A policy frame. As a result, one of the first tasks of the evaluation was to identify the portfolio of E&A projects, which would form the subject of the evaluation. Once we had done this, we analysed the portfolio, the findings from which are presented in the portfolio synopsis.

The project set analysis is the core of the macro evaluation, where we test hypotheses relevant to each project set to generate evidence of what works, when, where and why.

This methodology note explains in detail the process we have undertaken to identify the E&A project portfolio and the approach applied in a pilot analysis of a small set of social accountability projects. The latter has helped the evaluation team to understand how to improve the project set methodology to achieve more robust results. An adapted version of this methodology will be applied in the forthcoming full round of project set analysis.

Figure 1: Macro evaluation components and methods



A.2 Portfolio synopsis: the tabulated mapping protocol

During the inception phase, the evaluation team screened all DFID projects in order to identify projects relevant to the E&A policy frame, and to form a sample frame for the selection of project sets for the macro evaluation. Once the projects had been identified, they were organised in a tabulated mapping database that we refer to as the TM – a searchable database of data on policy-relevant projects. The TM has a dual purpose:

- to facilitate the E&A portfolio analysis; and
- to make information on DFID's E&A portfolio publicly available.

Table 1: E&A policy relevance screening guide summarised

E&A element	Project intervention	Project example
1. Social (short route) accountability through increased engagement between service users (demand side) and service providers (supply side) **Premise: voice, choice and accountability in service delivery will improve the quality, accessibility and reliability of services and secure longer-term improvements in well-being**		
1a) Strengthening vertical social accountability (citizen oversight) around service delivery	Raising citizen awareness around rights/entitlements to budget allocations and services Supporting citizen monitoring/oversight mechanisms of local budgets and service delivery Media capacity building/ freedoms linked to social accountability	Accountability in Tanzania Programme (AcT) (Output 1, 2, 3) Protection of Basic Services Programme Phase III, Ethiopia
1b) Vertical responsiveness (to citizens) around service delivery	Removing barriers and improving direct access to decision makers, e.g. through user-service provider platforms Building awareness, capacity and incentives to respond to citizens around budget and service delivery obligations	Accountability in Tanzania Programme (AcT) (Output 4)
2. Political (long route) accountability through citizen voice and engagement in political processes and policy cycles **Premise: more inclusive and accountable political systems result in more progressive and better sustained policy impacts**		
2a) Strengthening citizen political participation	Citizen participation in electoral processes Increasing representation of excluded groups in positions of authority Supporting policy advocacy by issue-based coalitions of interest Supporting independent oversight of policy Community sensitisation and mobilisation Public awareness campaigns	Strengthening Political Participation in Bangladesh Democratic Governance Facility- Deepening Democracy Phase II component, Uganda (Output 2) Vietnam VEAP
2b) Strengthening political accountability and responsiveness (through legislative, judicial, executive and civil society redress/oversight mechanisms))	Strengthening judicial institutions Strengthening public audit function Strengthening parliamentary committees Strengthening Ombudspersons/ Human Rights Commissions Changing incentives to improve policy implementation Supporting performance measures and review of policy implementation Supporting systematic and transparent budget and policy processes (including through decentralisation) Supporting public policy consultation mechanisms	Strengthening Democracy and Accountability in the Democratic Republic of Congo (Output 1) Strengthening Public Expenditure Management, Bangladesh

E&A element	Project intervention	Project example
3. Economic empowerment through lowering barriers to accessing markets and jobs **Premise: sustained growth and poverty reduction combines an enabling environment for ’market accountability’ in public policy with direct support to individuals, groups and businesses to claim their economic entitlements.**		
3a) Strengthening the enabling environment for market accountability formal legislation and policies to ensure economic rights and enforce contracts, policies that additionally improve the climate for foreign and private investment and for regional/international trade, and which invest in the infrastructure and information gaps that connect poor and rural communities to markets and value chains	Supporting economic reform discussions as part of more inclusive policy processes (via project interventions as listed under 2a) Support to consumer charters and regulatory/ legal frameworks Supporting contract enforcement and open/ordered market competition	Rwanda Land Tenure Regularisation [200284] 25.4m Madhya Pradesh Rural Livelihoods Project – Phase ll [113617]
3b) Support to economic empowerment through claiming economic entitlements	Building individual and collective economic rights awareness and economic literacy Support to collective action (e.g. economic movements, unions)	Creating Opportunities for the Poor and Excluded in Bangladesh (COPE) [202958] Oxfam’s Ngorongoro Land Rights Campaign sub project grant under AcT Tanzania [200498]

The identification of E&A policy-relevant projects followed two phases:

(A) Phase 1: Project screening

The first step to identify E&A policy-relevant projects was to develop clear and actionable **inclusion/exclusion criteria**, and to operationalise them in detailed guidance materials. The screening process followed three steps:

1. Screening for policy relevance
2. Screening against basic project criteria
3. Screening for quality of evaluative material

Step 1: The checklist for the E&A policy frame specifies in detail the elements that a project must include to be considered policy relevant. To draw these boundaries, we followed DFID's conceptualisation of E&A with the three lenses of: social accountability; political accountability; and economic empowerment. Guidance materials were developed, tested and refined for use by our research assistants (see Table 1). Spot-checks and double-blind testing were applied to ensure consistent and reliable categorisation.

Step 2: In the second step we applied some basic inclusion/exclusion criteria that had been agreed with DFID colleagues during the inception phase, presented here in Table 2.

Table 2: Inclusion/exclusion criteria

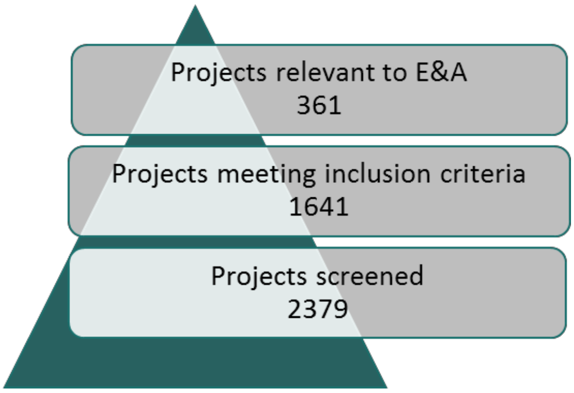
Inclusion/exclusion criteria	Justification
Total budget of £500,000 or over	This criterion serves to exclude most contracts for design, research, evaluation and other analytical work, which do not engage in change processes on the ground; 131 projects were excluded based on this criterion.
Based in one of DFID's 33 focal countries (including regional programmes in Africa, Asia, Central Asia and the Caribbean), GPAF projects, or Programme Partnership Arrangements (PPAs)	This follows the ToR, which indicated a focus on DFID's bilateral programmes, GPAF projects, and PPAs. The list of DFID's bilateral programmes as of 2011, when the two policy frames were adopted, was used. We did not search other countries, but among the projects we had identified through other means (keyword searches, gender marker, etc.) we excluded 27 based on this criterion.
Start date after 1 January 2011 or end date after 1 January 2013	This criterion serves to include projects that are likely to have been informed by the two policy frames. Both policy frames were launched in 2011. We have also decided to include older projects that have been running for at least two years since the launch of the policy frames, assuming that these projects are also likely to have been influenced by the policy commitments. This approach has helped to increase the project population while still maintaining a focus on the period after the new policy commitments; 448 projects were excluded based on this criterion.
No research projects, humanitarian interventions or evaluations	In addition to the £500,000 criterion, we found that there are a few research projects or evaluations with larger budgets. We have recorded these projects for learning purposes, but will not include them in the TM, since they do not engage in change processes on the ground; 25 projects were excluded based on this criterion. Humanitarian interventions were excluded, since they are typically short term and do not engage with empowering change processes; 63 projects were excluded based on this criterion.

No core funding to organisations (with the exception of the PPAs)	During the process, we found that there are a number of projects that provide core funding to organisations. Typically, project documentation for these projects does not include sufficient information on change processes on the ground and DFID investments are hard to trace. This exclusion criterion was at times fuzzy, with the project documentation not always being clear enough; 44 projects were excluded based on this criterion, many after discussions with the team.

The logframes sourced from DFID's Development Tracker website[1]were used to screen the projects against these criteria. All decisions were systematically recorded in a spreadsheet, which spells out the conditions that have to be met for a project to be included. Where there was uncertainty about how to code a specific condition on a project, this was recorded and later resolved in discussions with the wider evaluation team.

In total, 2,379 projects[2] were screened and 1,641 were identified as meeting the inclusion criteria above. This comprised 361 projects identified as policy relevant to empowerment and accountability according to the screening criteria presented in Table 1. The figure below presents this process graphically.

Figure 2: Project selection process



Step 3: The policy-relevant project population was further subjected to a screening based on the evaluative data quality of each project in order to identify a projects sampling frame for the project set analysis (see sampling protocol discussion below). This was based on an check on the availability of evaluations, or of reviews of the quality of the best available evaluative document, conducted by research assistants, with a 4-point graded Likert scale score applied to this document based on three quality criteria of**: (**1) whether there was a triangulation of data sources; (2) whether there was a degree of transparency in the assessment; and (3) whether there was a trustworthy analysis of the projects' contribution to outcomes.

From this quality screening process, a project set sample frame of 48 projects with sufficient quality evaluative data emerged.

Project screening challenges and lessons

Some challenges and lessons emerged while completing this activity. The boundaries of the policy area were not always clear and needed to be refined in an iterative manner. Recurrent themes of discussion included, for example, guidance on when to include higher-level governance reform projects in the E&A policy area. It was decided that higher-level governance reforms that at least open up opportunities for more accountability would be included; e.g. through opening up budgets, introducing citizen participation or monitoring,

[1] http://devtracker.dfid.gov.uk/
[2] All DFID projects in DFID priority countries, found in DevTracker and QUEST.

strengthening horizontal oversight mechanisms through other branches of power such as the judiciary, parliament, anti-corruption commissions or similar, etc. Conversely, public financial management projects were only included if they had explicit elements of transparency and oversight/scrutiny (horizontal or vertical), rather than only focusing on improving efficiency or budget allocations. Fuzzy boundaries were clarified in discussions with the core evaluation team, and all questions and answers were recorded in detailed guidance material for the research assistants. At times, this iterative refinement of the policy boundaries required projects to be reassessed.

Additional quality assurance processes included double-blind screening of a sample and random spot-checks. The double blind screening was achieved by a process where research assistants were independently assigned the same projects for the first two weeks of the task. Their decisions on the inclusion/exclusion criteria were then systematically compared and any differences resolved in a transparent discussion. This helped not only to quality assure the screening, but also to build a common understanding among our research assistants on how to interpret and operationalise the inclusion/exclusion criteria.

Random spot-checks where conducted by the evaluation team. The work of each research assistant as checked and some common errors in deciding to which lens the E&A projects were relevant were identified. On the basis of this spot-checking, the guidance note was tightened and examples of common errors were provided. The research assistants then undertook an additional quality assurance step and checked the coding of the E&A lenses for all projects identified as policy relevant to empowerment and accountability.

While working with a large team and applying qualitative definitions will always leave room for subjective interpretation, the above processes demonstrate how we have reduced the margin of error.

(B) Phase 2: Constructing and populating the tabulated mapping

This section presents the process of constructing and populating the database. The process of constructing the TM involved setting up coding categories in the software; downloading project documents for all projects identified as policy relevant to E&A from DFID’s internal management information system QUEST and the external platform DevTracker, and uploading onto EPPI-Reviewer;[3] reviewing project documentation; and coding the projects. For each policy-relevant project, logframes, business cases, annual reviews, mid-term reviews, project completion reviews and evaluations are stored on the TM (where available).

Two sets of codes were used to classify projects. The first set related to descriptive project data from which the E&A portfolio synopsis is derived. These include codes such as “start date”, “country” or “budget” of a project and are set out in Table 3 below.

The second was an evaluative set of coding of data quality of each project to identify projects to be considered for the project set analysis, as each project set requires a minimum data quality to be useful. This quality assessment included an appraisal of the availability of evaluations, or of reviews conducted by external consultants, and a 4-point Likert scale scored assessment of evaluative data quality based on three quality criteria of triangulation, contribution and transparency. The full details for this are presented in Table 3 below.

[3] EPPI Reviewer 4 is the systematic review software that hosts the database.

Table 3: Data in the tabulated mapping

Basic project characteristics	Value range
Start date	2009 (or before) to 2014
End date	2013–2017 (or later)
Duration	1–10 years (or more)
Geographical location	DFID's 33 focal countries (including regional programmes in Africa, Asia, Central Asia and the Caribbean)
Total project budget	£500,000 to £100 million (or more)
DFID's contribution to the project's budget	£500,000 to £100 million (or more)
Overall relevance or component relevance	Overall relevance to E&A, component relevance to E&A
Relevant component budget	£500,000 to £100 million (or more)
E&A lens	Social accountability, political accountability, or economic empowerment
Project description	*Narrative text*
Latest outcome score	A++, A+, A, B, or C
Quality of project data	
Available review documents	Annual reviews/ARs, mid-term reviews, project completion reviews, or evaluations
Number of available annual reviews/reports	0–3 (or more)
External or internal authorship of the 'strongest' review document	Internal or external
Planned evaluation	Yes or no
Degree of triangulation in the strongest review document	Strongly agree, agree, disagree or strongly disagree
Degree of contribution analysis in the strongest review document	Strongly agree, agree, disagree or strongly disagree
Degree of transparency in the strongest review document[4]	Strongly agree, agree, disagree or strongly disagree

As with the task described above of identifying policy-relevant projects, comprehensive quality assurance processes, including piloting, double-blind coding of a sample of projects and random spot-checks, were applied to this TM coding process. For the first two weeks of the task, research assistants were independently assigned the same projects. Their coding decisions were then systematically compared and any differences resolved in a transparent discussion, feeding back to the wider team. The TM software facilitated this process and helped the research assistants to build a common understanding on how to code the projects and how to assess data quality. A second round of double-blind coding was conducted a few weeks into the task to assure the quality of progress to date. Random spot-checks for each research assistant by the evaluation team complemented this approach. Overall, we found that there were some divergent opinions on some of the codes, but the margin of error remained manageable. Less than 10% of projects were found to be coded incorrectly when checked by the evaluation team, and this percentage was further reduced through the checking process.

TM challenges and lessons

The iterative nature of the process was again one of the key challenges and lessons. Coding criteria had to be refined and, at times, it was necessary for research assistants to revisit projects once new guidance was provided. For instance, some research assistants were coding average output ratings when outcome ratings were not available, thereby skewing the

[4] Note that the last three quality questions are based on adjusted and simplified quality criteria from the Bond Evidence Principles and DFID SEQAS quality standards for evaluation.

data by including two different types of data within the same code. This was resolved and defined in more detailed guidance, allowing us to ensure consistency as much as possible. Some of the coding criteria proved to be more subjective than others, in particular the three coding criteria used to assess the quality of data. We therefore combined this quality assessment with more objective quality criteria such as the availability of evaluations and externally conducted reviews.

Finally, it proved much more difficult and time-consuming to access QUEST and download the relevant project documentation than originally envisaged, resulting in an extended timeline for the population of the TM. Project documentation on QUEST is not organised in a way suitable for this task, requiring our research assistant to manually scroll through all available documentation and identify relevant documents. For some projects, this implied screening hundreds of documents, requiring a significant amount of time. We were not able to speed up this process by increasing the number of RAs because we were only able to obtain security clearance for one of our research assistants within the timeframe of the construction of the TM.

A.3 Project set analysis methodology

The realist synthesis approach

The social accountability project set analysis phase of the macro evaluation (see Figure 1) was based on the realist synthesis approach[5] that takes a holistic view of how interventions take place and what constitutes success. This approach synthesises a wide range of evidence to identify underlying causal mechanisms and explore how they work under what conditions. It seeks to answer the question ‘What works for whom under what circumstances?’ rather than simply ‘What works?’ only.

Contrary to experimental and quasi-experimental approaches in evaluation design, the realist synthesis approach assumes that detailed knowledge of context matters, as well as detailed, practical knowledge of how interventions have been managed. The realist synthesis approach is thus geared to produce knowledge that is practically useful and applicable in a variety of contexts.

Realist review is still a relatively new strategy for synthesising research, which has an explanatory rather than judgemental focus. Specifically, it seeks to ‘unpack the mechanism’ of how complex programmes work (or why they fail) in particular contexts and settings, and it is thus particularly well suited to an evaluation that has a strong focus on learning. A realist synthesis follows similar stages to a traditional systematic review,[6] but with some notable differences:[7]

- The focus of the synthesis is derived from a negotiation between stakeholders and reviewers, and therefore the extent of stakeholder involvement throughout the process is high.
- The search and appraisal of evidence is purposive and theoretically driven with the aim of refining theory.

[5] See Pawson, R., Greenhalgh, T., Harvey, G. and Walshe, K. (2004),‘Realist synthesis: an introduction’, *RMP Methods Paper* 2/2004. Manchester, UK: ESRC Research Methods Programme, University of Manchester.

[6] ibid.

[7] See Rycroft-Malone, J., McCormack, B., Hutchinson, A., deCorby, A., Bucknall, T., Kent, B., Schultz, A., Snelgrove-Clarke, A., Stetler, C., Titler, M., Wallin, L. and Wilson, V. (2012). ‘Realist synthesis: illustrating the method for implementation research’, *Implementation Science*, 7:33.

- Multiple types of information and evidence can be included.
- The process is iterative.
- The findings from the synthesis focus on explaining to the reader why the intervention works (or why it does not) and in what ways, to enable informed choices about further use and/or research.

To this end we operated in 'grounded theory'[8] mode. In contrast to a positivist research paradigm, grounded theory builds and tests working hypotheses that emerge from data as they are extracted, coded and synthesised, rather than in deductive mode in which a complete set of a priori hypothesis are confirmed or refuted.

Sampling protocol

We developed a simple two-step sampling protocol for the macro evaluations. In the case of the Social accountability pilot project set analysis this proceeded as follows:

Step 1: The project set sampling frame of 48 projects was derived from the three-stage coding exercise described above that screened from the total E&A portfolio projects that: (a) were relevant to the social accountability lens; (b) fulfilled the inclusion criteria (see Table 1); and (c) that achieved quality scores of 4 ('strongly agree') or 3 ('agree') against the three quality criteria of triangulation, contribution and transparency.

Step 2: Having confirmed that this social accountability project set of 48 projects shared an outcome of improved service delivery (see Qualitative Comparative Analysis (QCA) method below) we purposively sampled 15 projects in order to capture total project variability and to ensure coverage of the range of social accountability 'operational models' that we identified from our initial review of the project portfolio (see Table 1 and discussion in the main document).

While we are confident this sample size captures the variability in this project set well, with more time we would be able to expand the sample size and incorporate up to 100% of the total project set, which would allow us to increase confidence of inference from the QCA synthetic configuration findings while also increasing the scope analytical narrative discussion.

Data synthesis and analysis

Working with this sample project set, we employed two methods to synthesise and analyse the evaluative data. The rationale for using two methods was to (at least partially) bridge the gap between case-based analysis and 'large n' studies by quantifying qualitative elements of complex projects, while also retaining the in-depth interpretive analysis that can explain contribution to change in those projects. **First**, using the **QCA** method we first systematised the range of 'conditions' (comprising contexts, mechanisms and outcomes) relevant to the project set and applied a binary score (1=largely present; 0=largely absent) to each condition for each project in the project set. The resulting data sheet then allowed us to identify patterns, or 'synthetic configurations' of conditions that would give rise to the given outcome.

QCA is a case-oriented comparative approach that combines in-depth case studies with the identification and interpretation of causal patterns (Befani 2013[9]). QCA was first described by Charles Ragin[10] in the late 1980s as a method that sought to bring together the best features

[8] See Mills, J., Bonner, A. and Francis, K. (2006). 'The development of constructivist grounded theory'. *International Journal of Qualitative Methods*, 5: 25–35

[9] Befani, B. (2013) "Between complexity and generalization: Addressing evaluation challenges with QCA" *Evaluation*, vol. 19 no. 3 pp. 269-283

[10] Ragin, C. (1987). The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies.

of case-based and variable-based methods, or of qualitative and quantitative approaches.[11] The QCA approach enables the systematic comparison of cases, with each case viewed holistically as a complex configuration. A configuration is a specific combination of factors, known as conditions, which produce a given outcome. QCA thus views outcomes as products of combinations of factors rather than individual factors, recognising that causality can be non-linear and complex, involving several contributing factors for an outcome to be achieved. This is in line with our realist approach, which suggests that successful E&A outcomes are likely to be the result of a number of context and mechanism factors working together as identified in our hypotheses.

However, formal QCA is not viewed here as an end in itself: 'rather, it is a tool to enhance our comparative knowledge about cases in 'small and intermediate-n' research designs'.[12]We therefore viewed QCA as a first systematic step of robustly comparing different combinations of contributing factors that lead to an outcome – including key contextual factors. This enables an analysis of headline findings that lends itself to further interpretation. **Second**, using these QCA findings around configurations of conditions we then sought to interpret and illustrate these patterns based on **narrative analysis**: a deeper comparative qualitative analysis of the evaluative material available. There was a degree of iteration in the early stages of this process as narrative analysis threw up additional conditions that were then QCA-coded across the entire project set.

Our use of narrative analysis draws on the principles of 'meta ethnography'[13] to synthesise, creating a whole bigger than the sum of its constituent parts, and which builds comparative understanding. It complements and supports the QCA dimensions of the methodology by seeking, though systematic approaches, 'to reveal similarities and discrepancies among accounts of a particular phenomenon'.[14]Narrative analysis provided the interpretive layer to support the formal QCA analysis, enabling further deepening and testing of the Context Mechanism and Outcome (CMO) configurations arising[15].

Finally, and as Figure 1 illustrates, the use of the QCA and narrative analysis methods was iterative. The narrative analysis threw up new relevant variables that were integrated into the QCA, coded for all projects and included in a new round of synthetic configuration analysis.

Limitations to the project set analysis methodology

Principal amongst the methodological limitations to this macro evaluation was that the data came from secondary sources, with variability emerging in the quality, coverage and analytical depth of that data. This limitation was a key risk considered in DFID's Evaluability Assessment[16] which concluded that there was secondary reporting data of sufficient quality to proceed on this basis. During the inception phase, the evaluation team screened data quality using three-fold criteria (as described above). While this confirmed a sufficiently large project sampling frame, it nonetheless limited our selection to certain projects in a way that may have introduced bias towards those projects that had been evaluated for underlying reasons (e.g. if they were seen as good practice for promotional reasons or if they were more politically sensitive than other projects). We assumed, however, that the choice of evaluated projects

[11] Rihoux, B. and Ragin, C. (2009): Configurational Comparative Methods. Qualitative Comparative Analysis (QCA) and Related Techniques, London, SAGE.

[12]Ibid.

[13]Noblit G.W. and Hare, R.D. (1988). *Meta-ethnography: Synthesizing Qualitative Studies*, London: SAGE.

[14]Barnett-Page, E. and Thomas, J. (2009).'Methods for the synthesis of qualitative research: a critical review', *BMC Medical Research Methodology*, 9: 59.

[15] For an early example of combination of QCA and Realist Evaluation, see Befani, B., Ledermann, S. and F. Sager (2007) "Realistic Evaluation and QCA: Conceptual Parallels and an Empirical Application", *Evaluation* April 2007 vol. 13 no. 2 pp. 171-192

[16]Davies, R., Marriott, S J., Gibson, S. and Haegeman, E. (2012) 'Evaluability Assessment for DFID's Empowerment and Accountability and Gender Teams', Bristol, IDLGroup, July

was sufficiently probability-based not to have introduced this bias. We were, in any case, able to draw on a large proportion of projects with evaluative content in project reporting via Annual Reviews, Mid-Term Reviews and Project Completion Reviews. Furthermore, a significant number of these regular reporting documents were independently conducted, either by external consultants or by DFID staff external to the team responsible for overseeing project implementation.

Second, we were also limited by the variation in life cycle of the projects in the portfolio, with only the 'maturing' projects generating evaluative evidence that include outcome level contribution analysis. This will be mitigated to some extent by the extended nature of this evaluation timeframe, meaning that projects in the portfolio will mature and new evaluative material will become available. This will allow us to build new projects incrementally into future project set analyses.

Third, we were limited by the relatively small size of our pilot project set. With a small project set, this placed constraints on the confidence with which we could stratify according to attributes, or 'conditions' within the project set. This type of stratification within the Social accountability project set is pursued, for example, to compare operational models (as categorised in Table 1 in the main report) or isolate the influence of a contextual condition[17] of interest (for example contexts with a weak 'social contract' between government and citizens) to test its effectiveness in achieving a range of outcomes.

This reflects both the relatively small size of the pilot project set (15 projects) and the trade-offs involved in the application of QCA to this kind of synthesised analysis approach. The great advantage of QCA is its suitability to inductive learning (as discussed above). This means that rather than testing hypotheses in a conventional experimental sense, QCA codes all 'cases' (projects) for a wide range of potentially significant 'conditions', which can then be examined in different configurations and tested for their significance in contributing to a given outcome. The methodological caveat here is related to the need for 'limited diversity' in our QCA analysis in order to make our findings meaningful.[18] In other words, if we include too many conditions in the 'synthetic configuration' then we are unable to include a sufficiently large number of projects with that configuration that will enable us to be confident about the robustness of the finding. This caveat becomes more crucial the smaller the project set we start with. In this case, with a relatively small project set of 15 cases, we quite quickly run into this problem, notably when looking at specific operational models within the project set, as described in Table 1.

Fourth, we acknowledge that binary codings of the presence or absence of certain contexts is crude, that context can differ greatly within projects as well as between them, and that context can change over the lifetime of a project. Nonetheless the QCA configurations provide a useful synthesising entry point to identify patterns/associations for further interpretation. In respect of a single outcome as the basis for clustering our project set, we similarly acknowledge the analytical limitations of a binary score of 'sufficient evidence of achievement' (score 1) or insufficient evidence of achievement (score 0) applied to complex realities. However, the QCA method is not an end by itself but a tool to ask more focused questions to understand qualitative associations and change processes. The role of QCA – in capturing complexity in conditions and in dealing with multiple cases -- is to help bridge the gap between individual case-based learning and 'large n' studies so that we can make associational claims with more

[17] As discussed in the main report, Social accountability interventions are heavily context dependent (O'Meally, 2013, op cit).

[18] See, for example Davies, R. (2014). 'The Challenges of using QCA', blog posted at http://mandenews.blogspot.co.uk/2014_03_01_archive.html

confidence but without making claims for external validity at scale while retaining a realist understanding of 'systems' of complex change.[19]

Finally, and linked to the above, we note that binary scoring of a single outcome with a small project set can limit diversity in a way that denies us the ability to make meaningful analysis of the contribution to change of different combinations of conditions. Notably in the case of Social accountability there was almost always some evidence of improvements in service delivery that could justify a score of '1'. In order to achieve greater nuance and show greater diversity in outcomes, we made a scoring distinction between short-term local improvements in service and 'higher level' longer term improvements, as discussed in the main report. However, this did not significantly increase diversity in outcomes and we were unable to sufficiently contrast successful with unsuccessful cases. Our QCA findings at outcome level are therefore to a large extent limited to describing successful cases rather than contrasting what works with what doesn't. We have therefore also introduced a number of 'intermediate outcomes' that enabled us to shorten the causal chain and identify greater variation in scores. These adaptations in the QCA that emerged from the iteration of the two methods produced the final set of QCA conditions.

Despite these constraints we were able to demonstrate the utility of the methodology by limiting the diversity of conditions considered in any single QCA round of analysis, while including intermediate outcomes in these rounds, and to sequence this analysis with in-depth interpretive narrative analysis as described above. In this way we do not make inflated claims for validity with either the QCA or narrative analysis approach, but instead harness the comparative advantage of both methods through iteration and demonstrate that this iteration allows us to identify meaningful associations and interpret those associations.[20] Hence these limitations should be considered in view of the major strengths of the approach and methods; namely that they offer a 'systems' view of context, mechanism and outcomes, and within the boundaries and definitions set for the task, provide an opportunity to generate evidence on what works, when and where, and why.

Finally, the social accountability pilot was conducted under time constraints that reduced the opportunity for further refining and recalibrating QCA conditions, introducing quality assurance ratings, expanding the number of rounds of QCA configuration analysis and producing a more comprehensive check back of findings to existing literature on what works in Social accountability. In the presence of a high number of conditions and cases, the usefulness of QCA is proportional to the time available to explore the dataset, spot and interpret patterns. In this case there has been a considerable distance between the variety of patterns spotted in the inductive phase, and the time available to make sense of them, which has made the team focus on the hypothesis testing QCA phase. Nonetheless, this pilot methodology provides a platform for considering expanding this project set to accommodate a larger number of projects from across all three E&A lenses (social accountability, political accountability and economic empowerment). It will be an additional opportunity to test the hypotheses which have emerged, to interpret the patterns identified, and – when the same number of conditions is kept – it will provide stronger assurance that the patterns emerged from the Boolean minimisations are not due to chance[21].

[19] See for example, White, H. and Phillips, D. (2012, 8). 'Addressing attribution of cause and effect in small n impact evaluations: towards an integrated framework', International Initiative for Impact Evaluation, *Working Paper* 15, Delhi, 3ie, June.

[20] For readers with a more technical understanding of QCA, it is also important to note that the lack of diversity in outcomes affects most strongly the opportunity to conduct the INUS analysis (i.e. when you compare two combinations which are identical except in one condition BUT they have a different outcome), while the validity of the findings of both the Boolean minimisation and the subset-sufficiency analysis of the frequently present outcome might still hold to some extent, even without a large number of negative cases.See Befani, B. (2013) "Between complexity and generalization: Addressing evaluation challenges with QCA" *Evaluation*, vol. 19 no. 3 pp. 269-283 and http://eba.se/en/evaluating-development-interventions-with-qca-potential-and-pitfalls/, forthcoming

[21] See http://eba.se/en/evaluating-development-interventions-with-qca-potential-and-pitfalls/ forthcoming