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

Tools and Guidelines for Improving the Evaluability of INGO Empowerment and Accountability Programmes

Abstract This Practice Paper Annex is the result of an analysis of INGO evaluation practice in empowerment and accountability (E&A) programmes commissioned by CARE UK, Christian Aid, Plan UK and World Vision UK. It is the companion to CDI Practice Paper 01 that considers the implications of current evaluation and learning debates for those seeking to improve the quality of evidence pertaining to the outcomes and impacts of INGO empowerment and accountability programmes. This paper proceeds from the premise that if international NGOs are to successfully 'measure' or assess the outcomes and impacts of E&A programmes, they need to shift their attention from data collection tools to a more holistic approach to evaluation design. Strategies need to be appropriate given organisational values, evaluation objectives and programme attributes, which include programme contexts. The authors present a series of practical tools for use and adaptation by INGO staff, donor representatives and consultants keen to improve the evaluability of E&A programmes.

Introduction

This Practice Paper Annex is the companion to CDI Practice Paper 01. Both papers are based on a review commissioned by CARE UK, Christian Aid, Plan UK and World Vision UK using funding from their DFID Programme Partnership Agreements (PPAs). We conducted a review of evaluation documents pertaining to 16 empowerment and accountability (E&A) programmes and projects implemented by the four international NGOs (INGOs) which deployed a wide range of different methodological designs and methods and had diverse purposes. CDI Practice Paper 01 considers the implications of current evaluation and learning debates for those seeking approaches to assess the outcomes and impacts of INGO E&A interventions, and produce 'quality evidence'. This paper presents tools that we developed in the course of reviewing, analysis and reflection:

- Tool 1: Evaluation design logic table
- Tool 2: Programme attributes analytical framework
- Tool 3: Guidelines for improving the evaluability of E&A programmes
- Tool 4: Glossary of terms used in contemporary MEL debates and documents

These tools are intended to assist development actors – researchers, evaluators, consultants, INGO and official aid agency staff – when designing INGO monitoring, evaluation and learning (MEL) systems capable of assessing the 'results', outcomes and impacts of empowerment and social accountability programmes and enabling learning from the process.

We start from the premise that if INGOs are to successfully 'measure' or assess outcomes and impacts of E&A programmes, they need to shift their attention from indicators and data collection tools to a more holistic approach to thinking about appropriate monitoring and evaluation strategies and systems. If they are to lead or commission evaluations and impact assessments that generate evidence of the desired quality, and are consistent with their organisations' participatory values, they need to start developing MEL strategies at the programme planning and budgeting stage.

A further starting assumption is that developing MEL strategies and systems that support such analysis and learning will inevitably be an iterative and imperfect process. It will need to be done differently according to the values of the different organisations implementing them; the relative strategic importance of programmes and consequent framing of evaluation objectives and questions; resources available for evaluation; specific programme attributes; and the contexts in which they are implemented.

Tool 1: Evaluation design logic table

While many debates and discussions are at the level of methods and kinds of data, contemporary MEL challenges and our review highlight the importance of differentiating between **methods** (approaches to data collection and measurement tools and statistical analysis) and **design** (the overarching logic for evaluations that includes evaluation questions, theory used to analyse data, data and use of data). Design logic needs to be internally consistent, so that the kind of data produced and the methods chosen to produce them are determined by the methodological design, and the methodological design is determined by the evaluation or research questions that need answering. A further consideration is the degree of strategic importance of the programme in question within the organisation or the overall thematic portfolio. Cost–benefit considerations mean some organisations are selective in their choice of programmes for in-depth evaluation or impact assessment for learning that has broader implications for their work.

Tool 1 can help to assess how strategic considerations influence choice of evaluation questions and appropriate evaluation design, and to orient those seeking to develop evaluation designs that respond to particular strategic concerns.

Table 1: Tool 1 – Evaluation design logic table

Evaluation objective/ strategic importance of programme	Type of question and assumptions	Nature of causal explanatory requirements and analysis	Design issues and comments in relation to INGO E&A programmes
a) To generate knowledge/ evidence that can be used for accountability to donors and taxpayers	- To what extent can specific outcomes and impact be attributed to an intervention? - What is the net effect, e.g. number or % of people experiencing x level of improvement? *Assumptions:* outcomes clearly understood; possible to isolate cause and effect; no interest in generalisation to other interventions	- Counterfactual – need to be able to manipulate the intervention and have large number of households or communities	- Experimental or quasi-experimental. - Few E&A programmes implemented by NGOs meet these requirements. This question can be impossible to answer.[1] - Hybrid designs including use of theory-based designs, case studies and/or participatory processes can shed light on questions.

[1] Mayne www.cgiar-ilac.org/files/publications/briefs/ILAC_Brief16_Contribution_Analysis.pdf

b) To demonstrate NGO effectiveness and accountability to donors.	- Has there been a change? What influenced the change? Is there reasonable evidence to suggest our programme has influenced it? Was the intervention vital for the effect? Was it sufficient or did other factors help? *Assumption*: Likely to be multiple causes responsible for any change observed	- Identification or confirmation of causal processes or factors supporting change in context -Identification of possible alternative explanations and confirmation that initiatives were not [or not solely] responsible for the change - Comparable cases where common set of causes are present and evidence of their potency identifiable	- Theory-based evaluation approach to exploring causal mechanisms And/or Case studies that explore causal links - Both of above approaches likely to include some relevant quantitative data, contextualised with appropriate analysis, e.g. process tracing of qualitative data and contribution analysis
c) Learning for policymakers and programme managers	- How and why did the change/outcomes identified in (b) happen? *Assumption*: interventions interact with other causal factors, but it is possible to plausibly suggest causal mechanisms	-Identifying relationships between programme and context - how the latter has influenced change. -Identification of causal mechanisms	- Theory-based, especially realist evaluation design that integrates context analysis using approaches with attributes mentioned above
d) Practitioner and citizen learning	- How and why are change/outcomes happening or not and can we do anything to enhance potential impact? What is the influence of changes in the environment? Are our initial assumptions correct? Are we doing the right things? What is the influence of the quality of our delivery mechanism? *Assumptions*: same as (c)	-Understanding relationship between programme and context - how the latter has influenced change. -Identification of causal mechanisms, validation of cause and effect by stakeholders.	- Theory-based, real-time evaluation to enable learning and adaptation during implementation with attributes mentioned above. Inclusion of participatory elements, especially feedback to participants to increase accountability to citizens as well as impact
e) Learning about whether similar programmes are likely to achieve similar outcomes elsewhere	- Transferability – is it likely that whatever happened here could happen elsewhere? - What generalisable lessons have been learned about outcomes and impacts? *Assumption:* [some of] what has enabled change or stopped things getting worse in one place can work elsewhere	-Identification of causal mechanisms, validation of cause - effect relationships by stakeholders. -Identification of factors that helped/hindered change developed into a typology with possible relevance elsewhere, e.g. the nature of civil society in certain countries may mean similar programmes are more likely to have similar effects than they would in entirely different contexts.	- Theories of change, participatory and, case study approaches, synthesis studies (e.g. see Christian Aid Governance and Transparency Fund (GTF) Mid-term Review by McGee and Scott-Villiers 2011 for application of such thinking)

(Adapted from Stern *et al.* 2012)

Tool 2: Programme attributes analytical framework

Systematic consideration of programme attributes and the contexts in which they are being implemented helps to identify specific evaluation challenges, the strength of causal inference possible, and implications for evaluation design. This tool – which draws heavily from Stern *et al.* (2012) – summarises some key attributes of E&A programmes, the evaluation challenges they pose, and their implications for evaluation design. It is intended for use within a holistic process described in Tool 3: Guidelines for improving the evaluability of INGO E&A programmes (below).

Table 2: Tool 2 – Programme attributes analytical framework

Attribute	Evaluation challenges	Implication for design
Nature of outcomes and impacts and how easy are they to measure/observe: Some E&A programmes that focus on service delivery appear to have easy to define outcomes, e.g. improvements in access to health or education services. Others concerned about empowerment and shifting power and accountability relationships between different actors face more challenges when assessing results. These, less tangible changes, are harder to identify and measure since different stakeholders have different perceptions of what key concepts mean.	- Impacts and outcomes have different meanings for different stakeholders and beneficiaries - Deciding what it is that is the most important thing to observe, assessed or measure	- Jointly develop a theory of change through a participatory approach with inputs from stakeholders and beneficiaries that help to define outcomes and impacts. Then decide together how to assess. - Alternatively, incorporate open-ended 'Most Significant Change'-type questions that can be translated into more standardised 'empowerment' indicators by management agencies for quantification. (This is an expensive option). - Pilot the use of simple power analysis mapping approaches to exploring power relations during context analysis/baseline to help identify power relations the programme wants to shift. Repeat at appropriate stages during programme implementation and final evaluation
Complex set of actors and relationships involved: Likely to be a particular concern for non-operational organisations that manage civil society funds and also work with partner NGOs who work with community-based organisations. Can be a challenge in programmes implemented in partnership with government agencies.	- Deciding how to develop a manageable evaluation strategy with clearly assigned responsibilities - Negotiating agreement and standardising data collection for analysis - Deciding appropriate levels of evaluation with accepted measures for civil society organisation capacity building	- Identify distinct stages. a) Develop and use a theory of change that shows how different actors contribute to different results; b) Devise a nested strategy that evaluates some components discretely from others. In complex programmes this could mean having distinct evaluation strategies for different 'outputs'. - Consider contracting technical support for evaluation design
Customised non-standard projects implemented in diverse contexts: This is a common problem experienced by INGOs using funds managed by institutional donor Headquarters, e.g. DFID's Governance and Transparency Fund and Programme Partnership Agreement funding managed by the London office.	- How to add up apples and oranges	- Identify alternative, generic theories of change (TOCs) (e.g. Jonathan Fox's framework used in the mid-term review of Christian Aid's GTF programme, by McGee and Scott-Villiers 2011) and use them as tool for consolidating results. - Focus on understanding mechanisms rather than effects - Develop context typologies to enable learning about what works better where and why - Involve stakeholders in participatory designs at local level

Time to achieve impact and nested programmes: Some E&A projects funded by donors are nested within long-term INGO strategies of work with broader populations. These can take the form of community-based initiatives or capacity development for civil society organisations contributing to national level advocacy	- Pressure to develop 'false' baselines within programmes that have effectively already begun and been subject to previous baseline exercises. - Untangling the effects of donor funding from a longer-term programme of work. - Being pushed to define impacts within the 'false constructs' of donor funding instead of the more appropriate local parameters of the longer-term programme.	- Explain to donors the challenges, including the risks of overambitious and ultimately cost-ineffective evaluations or impact assessments. - Develop nuanced qualitative TOC, participatory or case study hybrids likely to be more effective and meaningful than an approach that tries to quantify a net effect or impact. - Construct extended TOCs and use monitoring systems and indicators to assess distance travelled, decide when best to do evaluations and track critical events that allow testing of assumptions and redirection of programmes
Likely overlap with other programmes: Except when implementing directly in areas unaffected by activities of any other development actors or other programmes being implemented by the same agency, it is unlikely that INGO E&A programmes will be the only contribution to changes or outcomes of interest.	- Disentangling effects from other programmes - Disentangling contributions of different actors	Joint evaluations using contribution analysis (Mayne n.d.). Particularly relevant for programmes funded by same donor. When there is one institutional donor, may be worth discussing with donor the extent to which it sees the INGO's project contributing to its broader governance aims in a particular country, and to its longer theory of change. This might require donor to adopt more joined-up, nested, country level evaluation strategies.
Desire for spillover: Many sub-national E&A initiatives actively aspire to spill-over effects through self-replication that offer possibility of enhancing effectiveness	- Potential challenge if considering experimental designs	- Avoid experimental designs and seek ways to ensure potential spill-over areas are included in evaluation design. - Explore the nature and degree of self-replication/spread or lack thereof as a discrete aspect of evaluation
Non- linearity and emergent outcomes: Few E&A programmes have linear properties. The quantity of inputs does not have a direct relationship to the quantity of outcomes. Complexity and systems theory better describe unpredictable pathways of change – empowerment of marginalised groups may lead to pushback from vested interests before they lead to improvements	- Appropriate timing of evaluation activities - Specific measures may not be available in advance, which makes longitudinal studies difficult	- Real-time/realist evaluation approaches that explore change processes in order to explain 'resistance' etc constantly review and adjust theories and expectations of results according to evolution of the context
New approaches: E&A work is essentially political and programmes often attempt innovative approaches	- Programme designers can have problems trying to develop programme theories underpinned by verifiable assumptions	- Participatory and/or theory-based design - Real-time evaluations with constant exploration of alternative causal mechanisms and pathways to change.

Context, uncertainty and risk: The nature and degree of change possible is context-dependent, not only at national level, but also at district and commune levels. In post-conflict situations possibilities for change can vary dramatically across communities. Contextual challenges to change also influence the ease of measuring it in different places.	- Likelihood of setbacks and uneven progress - Access to stakeholders and overcoming bias that can emerge from fear for personal safety etc. Can be a particular issue in participatory designs	- Evaluability assessment - Need for real-time or formative evaluation approaches that enable feedback to managers, not only about outcomes of programme, but also about feasibility and ethics of different monitoring and evaluation methods.

(Adapted from Stern *et al.* 2012: 60)

Tool 3: Guidelines for improving the evaluability of E&A programmes

Set out as a list of questions, these guidelines aim to prompt thinking about issues that need to be considered when designing and implementing monitoring evaluation and learning strategies. They are aimed at UK or country-based programme managers, MEL staff, and those responsible for writing proposals or commissioning evaluations and impact assessments for donor-funded programmes. In some instances the guidelines may suggest the need for external assistance in some or all stages of the development and implementation of MEL strategies. Choosing designs and methods, collecting data, storing and analysing data, applying learning and communicating results all require different kinds of skills.

When should MEL, impact assessment or evaluation strategy design begin?

Key message:

- Start early to ensure that you maximise opportunities for MEL strategies to contribute to real-time learning to improve impact and the value for money (VfM) of programmes, evaluations and impact assessment.

Why start early? There are effectiveness reasons, value for money arguments and an ethical imperative for using monitoring, evaluation, and learning to improve performance over a programme's lifetime. For this, MEL systems need to be included in programme design. Effective MEL strategies in E&A programmes should seek to amplify impact by:

- Exploring opportunities to translate theories of change into participatory and locally-owned theories of change or actions at community level. This tactic to enhance effectiveness is being tried by some parts of World Vision and Care.
- Consolidating monitoring data from local-level E&A activities to:
    - share back with communities to enhance horizontal alliances required to support broader collective action and use in 'vertical accountability' initiatives, e.g. to support advocacy to achieve impact at scale;
    - enable 'real-time' learning and adaptation.

Another reason for early consideration of MEL strategies is to avoid cost-ineffective impact assessment caused by a lack of basic monitoring data. All evaluations and impact assessment require process-monitoring data. Sometimes this is to help triangulate other data on outcomes – e.g. minutes and attendance lists of meetings can often be used to support interview data on increased engagement between citizens and state actors. In the past INGO E&A MEL plans have missed such opportunities; however they are beginning to remedy this oversight.

Example

World Vision UK's Influence and Engagement Matrix and School Scorecards being trialled by Plan UK in Sierra Leone, Malawi and Cambodia seek to take advantage of scorecard rating that tracks perceptions of changes over time. Both are anticipated to enhance evaluation and causal explanation in the future. Schoolchildren's rating of their schools' performance over time will enable them to look at longitudinal change to reflect on whether their action plans are making a difference. At the same time it will provide a database that could help inform external evaluators' judgements about the effectiveness of the scorecards as tools to encourage E&A.

What needs to be considered when deciding on MEL designs?

Key messages:

- INGO organisational values and norms favour participatory design elements
- The current state of evidence on E&A impact and the complex, unpredictable and long-term nature of E&A programmes mean experimental and quasi-experimental approaches have limited utility as single designs or hybrids
- Theory-based approaches guided by TOCs are the most appropriate dominant design. They can be complemented by participatory elements or case studies
- Design decisions should be influenced by how important it is to infer causality in any given case
- Considering particular programme attributes and the contexts in which they are being implemented helps to identify specific evaluation challenges and the strength of causal inference possible
- Resource and capability constraints place pragmatic limitations on scope of MEL strategies and need to be considered early on.

How are organisational values/norms dictated in evaluation policy or elsewhere likely to influence our design choices? MEL strategies are shaped by organisational norms and values. Possible questions to consider include:

- Do all programmes need to give equal emphasis to evaluation, or does/can/should the NGO operate a differentiated evaluation policy? What does this mean for the particular programme under consideration? Tool 1 (Table 1: Evaluation design logic table) can help to assess how strategic importance is likely to influence choice of evaluation questions and appropriate evaluation design.
- What is the desired level of participation of different stakeholders in the design and implementation of the evaluation strategy?
    - E&A programmes are based on values, and consistency requires that these values (for instance, participation, empowerment, inclusion, accountability) be built into the way the programmes are monitored, evaluated and learnt from. Questions asked might include: What role should NGO and government partners play in MEL design, implementation, analysis etc? Is it sufficient that citizens' voices are heard through interviews and/or focus groups? Should citizens have a bigger role to play in the MEL design and analysis? Should more resources be allocated for this?
    - What do individual organisations' ethical principles or guidelines suggest about the relative merit of different designs and approaches? What are the opportunity costs associated with different types of approaches for different types of participants, e.g. women? If experimental approaches are being considered, what are the costs for people in 'control' areas?

How do specific programme attributes influence design? E&A MEL designs are likely to be complicated by virtue of their inherently political attributes (McGee and Gaventa 2011; Kelly and Roche 2011) and will require multiple designs and methods. Starting to explore early on what the particular attributes of a programme

mean for evaluation designs can help to refine ideal design types to ensure that evaluation strategies are feasible. Mapping out a programme's theories of change and action and exploring the following questions will enable this:

- Nature of change: what change is the programme trying to effect? Is it likely to?
- Sphere of influence: what factors are within the programme's direct control, indirect control, outside of its influence?
- Complexity of relationships: who needs to do what when for changes to happen?
- Assumptions: what are our assumptions about how change happens illustrated in different programme strands or pathways?[2]
- Context: what is known about the existing context?

Context analysis to inform programmes and contribution analysis

An early attempt at teasing out a theory of change (TOC) should lead to a better understanding of what is known about the context. Understanding knowledge gaps can shift baseline exercises from technical monitoring and evaluation (M&E) processes for filling in baseline indicators to rich 'reality checks' that are vital for informing programme design and learning.

Key questions include:
- Situation analysis: what is the current situation of different groups?
- Power analysis: what are the formal and informal institutions, power relationships, structures, norms and values that shape people's lives?
- How does (social/political etc) change tend to happen in this place? What has prevented change from happening in the past?
- What 'secular trends' are at work in the broader environment that affect this place now: increase in internet penetration? Major post-conflict road building?
- What other actors, including donors and INGOs, are already working or planning to work in this place with this population? Doing what, aiming at what?
- Evaluation methodology: what are the population demographics, how heterogeneous is the population and how might that affect sample design and associated costs? What statistics or data exists at national, district and local level that may provide secondary data for triangulation? What criteria are used in public statistics to identify poor and very poor groups? Are they consistent with those our programme is using?

Having mapped out some key parameters of the programme, it should then be possible to look at the particular challenges created by its attributes and their implications for MEL design. Tool 2 – Programme attributes analytical framework – summarises some of the key attributes of E&A programmes, the evaluation challenges they pose and their possible implications for evaluation design.

[2] Questions influenced by Mayne J, Contribution Analysis http://www.cgiar-ilac.org/files/publications/briefs/ILAC_Brief16_Contribution_Analysis.pdf (accessed 13 March 2013)

What needs to be considered when operationalising a MEL strategy?

Key messages:

- Operationalising MEL designs requires choosing which outputs, milestones towards outcomes, and – if appropriate – impacts, will be identified and assessed
- Decisions involve tradeoffs between local utility and standardisation for donors
- Sample design must consider demographics and power relations
- Choices of measurement tools must be guided by programme attributes (Tool 2) and consideration of how tools and triangulation approaches are likely to enable interrogation of TOC assumptions
- Quality is often more important than quantity when thinking about indicators, sample designs and measurement tools
- Make sure qualitative approaches consider the costs of translation and analysis.

Once a potential design has been identified, it is necessary to develop and test the feasibility of more detailed operational plans. This requires returning to the theory of change and identifying more specific outputs and outcomes that will be monitored and measured/assessed at different points in time, as well as methods for collecting data. Although likely to require refinement during inception stage of programme implementation, it is helpful to consider the following during programme design.

What changes/outcomes is it appropriate to assess to explore 'mechanisms' for change at different points in a programme's lifetime? Many of these changes/outcomes will be driven by NGOs' existing theories of change and be consistent with those identified in the BOND effectiveness programme 'Improve' it frameworks.[3] However, consideration of the programme attribute framework (Tool 2) should help to shape decisions about which outcomes or impacts it makes sense to focus on when. Given a real-time learning intention, they need to be mapped onto the TOC to assess what data ideally can be collected at different times to advance a learning agenda.

What type of measurement of change is required and why? Developing systems that can establish that programmes make a difference is difficult; assessing precisely what and how much difference they make to different groups of people compared to other environmental factors is, some would say, virtually impossible (Mayne n.d.). The risk in E&A programmes is that what is unquantifiable becomes unimportant and qualitative changes that throw up patterns that are interesting but can't be measured get ignored. To avoid this, a modest provisional combination of quantitative and qualitative indicators should be identified. They need to make sense in terms of their scope to test causal mechanisms, but will require refining and adaptation through more participatory processes with stakeholders during the programme inception phase. There is a trade-off between standardising the use of indictors for aggregation by those reporting to donors, and adaptation for local utility. Ensuring comparability of a few simple quantitative or qualitative data across locations means using the same time periods and definitions for at least a couple of indicators in all sites. Defining a few standard measures does not preclude the use of more context-specific indicators as well.

A 'fit for purpose' set of empowerment indicators is one which provides sufficient description of changes in power relations to frame and prompt in-depth analysis of those changes in ways that will lead to improved empowerment interventions and help hold decision-makers accountable for the impacts they have on people's lives (Holland 2010).

[3] www.bond.org.uk/data/files/Effectiveness_Programme/IIF_thematic_papers/Empowerment.pdf
www.bond.org.uk/data/files/Effectiveness_Programme/IIF_thematic_papers/GovernanceAccountability.pdf
(both accessed 21 March 2013)

Who is our population of interest and what are the key units of analysis? Defining a population implies understanding the elements that form the population of interest. It may consist of all individuals living in an area, or individuals with certain characteristics. Many studies will involve several units of analysis. For example most E&A programmes are interested in collecting data about groups or organisations, as well as individuals. Large studies may be interested in performance in particular geographical areas distinguished by physical features or administration units. More than one unit of analysis may be needed, e.g.:

- Hierarchies of units of analysis, which may be influenced by the articulation of indicators in logframes, e.g. x number of communities in y districts will require a sample of villages for each district to be reported on. For E&A programmes opportunities to triangulate data from different stakeholders at different levels of this hierarchy should be exploited.
- Horizontal strata that are commonly thought of as being at the same level but grouped according to specific characteristics have an effect on the issues of research interest. A nuanced power lens is likely to reveal groupings that will have a bearing on results e.g.:

    - Possible criteria to group villages: proximity to roads, main incomes of livelihood options, religions, political affiliations of local leaders etc,
    - Possible criteria to group individuals: Men, women, age, class, disability, ethnicity, religion, levels of education, sexual orientation. Possibilities to compare/'triangulate' perceptions of different groups should be explored.

What should the sample's nature, composition and size be for different outputs/outcomes? Every MEL strategy, no matter how simple, involves some degree of 'sampling'. However, the relevance, complexity and nature of sample designs vary significantly depending on the emphasis of the overall MEL design, programme attributes, population demographics and target groups.[4]

Unfortunately there is no magic formula to determine the correct sample size, and statisticians argue it will depend on the particular context. As a general rule of thumb, sample design complexity increases the more heterogeneous the general population is and the more criteria you use to define your particular target group. Three main factors influence decisions. The principles are applicable to qualitative and quantitative studies:

- *Variety* in the population with respect to the characteristic or issue of interest – the more varied, the larger the sample needed.
- *Levels of aggregation*. If reporting change at district level, it will be necessary to sample several villages within each district. In the district level design the number of villages sampled per district is likely to be larger than if the design aimed to report change at a provincial level, which would require a smaller sample of villages per district.
- *Resources*. If resources are tight it may not be possible to visit many villages. A compromise would be to invite people from different villages to a central location. With credible key informants who have the relevant knowledge and 'unbiased' views, this may be a useful approach. However such a strategy may have limitations if power relations are such that key informants present a biased view.

Tip

For designing a formal sample the Services Centre at The University of Reading advocates the employment of a statistician – adopting pre-packaged solutions or standard plans without thought is likely to be 'a recipe only for disaster!' (SSC 2001).

The above factors need to be considered carefully when deciding whether to build MEL strategies that focus on depth rather than breadth. It might be more cost-effective to randomly or purposively select a few case studies and undertake regular in-depth activities with groups of interest than risk expensive ineffective larger-scale sampling. Formal random sampling for generalisation does not always require a large number of sites, but if it is not possible or required, longitudinal comparison of case studies – purposively selected because they

[4] More user-friendly information about sampling and other evaluation tasks can be found on http://betterevaluation.org/plan/describe/sample (accessed 13 March 2013)

exhibit characteristics of particular interest – may be a robust alternative. Given the sheer impossibility of ever being able to quantitatively aggregate the effects or impacts of all programmes at an organisational level, NGOs need to decide what they can reasonably know or say about 'big numbers.' Going forward we might expect to see more articulations of results in terms being used by Oxfam GB. At an organisational level Oxfam GB reports quantitative change for a small sample of project case studies that have been rigorously evaluated, together with a statistic on the number of people estimated to have participated in its projects globally.

What 'research encounters' and data collection methods? Many empowerment and accountability programmes have tended to combine group activities – focus group discussions and activities using more participatory tools – with occasional use of individual or household-level interviews. In the past there has been a tendency for NGOs to extrapolate data gathered through focus group discussions and estimate effects on numbers of people living in areas. As pressure to improve the quality of evidence grows, NGOs are going to have to take more care and make more explicit the assumptions they are using to generate 'big numbers'. They probably need to reflect on whether reporting results in terms of numbers of individual people is consistent with the models of change that underpin E&A work, and also, with the sociocultural contexts in question (in some cultures individualism as a social value is far less prevalent than in Western cultures and collective values prevail). Results of much E&A work, especially that informed by more progressive ideas about citizen-led accountability initiatives, might be better reported on at household, group or community level.

Consistency between units of analysis. When designing MEL systems ask:

- Do the data collection methods proposed match the units of analysis used for articulation of outcomes and indicators? For example, if an indicator is expressed in terms of numbers or % of people, individual interview is the appropriate data collection method as opposed to household surveys or focus groups.

What data collection and measurement tools? Having decided on units of analysis and the types of research encounters, there is a need to select data collection methods to measure change. In recent months NGOs have responded to pressure to demonstrate results through developing a number of new tools to help in assessing change and results at different points in the results chain. For those developing approaches that not only assess change, but also NGOs' contributions to such change, the challenge becomes one of choosing tools appropriate for the overall evaluation design.

Examples: Testing of measurement tools and lessons learned by the 4 INGOs

Plan International has developed two scalar tools: one based on very participatory principles enables schoolchildren to identify issues they think are important in their schools, develop action plans to address them and then monitor progress in those areas over time. One limitation appears to be that it cannot measure things getting worse than they were at the 'baseline', if the baseline score is zero. The second tool, The Girl's Opportunity Star, that uses a scalar approach to encourage groups of girls to score empowerment in several dimensions chosen by Plan International, is also currently being tested.

World Vision has developed a standard tool for measuring engagement between communities and governments that is being trialled with advocacy targets (mentioned earlier). We note that it needs to be linked with other monitoring data to interrogate theories of change with appropriate contribution analysis.

Christian Aid's piloting of a perception-based rights-claiming tool illustrated the need to ensure tools applied at different points in programmes use exactly the same measurement approach; that rationales for choices of respondents need to be clear; and that in small-n studies the same individuals need to participate in baseline and longitudinal data collection exercises. It also generated useful lessons about the challenges of comparing data gathered using the same tool from very diverse contexts.

Care Bangladesh has piloted the Most Significant Change (MSC) approach as part of a participatory impact assessment design. Answers to open-ended MSC questions were coded by staff to fit with donor indicators before analysis. This draws attention to the possibilities of quantitative analysis from MSC, but also to the need for deeper analysis of what results mean in terms of causal connections. The consultant who managed the test identified a number of challenges. They included: the risks of manipulation by facilitators; developing appropriate sample designs, particularly as relates to hierarchical units of analysis; and time constraints related to managing large qualitative data sets.

The BOND 'Improve it' frameworks for assessing the effectiveness of empowerment, accountability and governance programmes include an index of tools for measuring change relating to common E&A outcomes and impacts.[5] Some appear to have been developed within the context of more detailed methodologies than others. The potential efficacy of each tool will ultimately be determined by how they are used within the specific confines of different programmes.

Many of the BOND Effectiveness Programme (BEP) scalar tools fit well with participatory and theory of change case study designs that take real-time approaches to learning and adaptive management. However, they are more challenging to use for absolute measurement in longitudinal studies. Providing they are used within small samples that allow those facilitating exercises to ensure that those involved in baseline and future MEL activities share enough characteristics to make assessment of change valid, they can be used for measurement. However, as different individuals can be involved and perceptions of values and ratings can change as participants increase their knowledge of criteria being discussed, they need to be handled with care. Some argue they are better used within a realistic framework that explores perceptions of change rather than longitudinal studies. Whatever the case, those that include efforts to triangulate within tools or overall designs, e.g. by getting external perspectives or linking them to more objectively verifiable indicators, are likely to contribute to more convincing stories of change.

Moreover, Likert scales use ordinal values and therefore cannot be subjected to parametric statistical tests that are commonly used in experimental and other statistical methods.[6] As the intervals between many of the concepts used in scoring ranges have no meaning, many statisticians argue that it is not possible to calculate a mean or a standard deviation. Averages should be calculated using medians and modes. The implications of this, and a more thorough analysis of the strengths and weaknesses of using scalar tools are included in a paper on scalar approaches by Jerry Adams (2012).

Although the review did not include consideration of participatory video or mobile phone technology, these are growing in popularity and World Vision has documented experience using mobile phones to collect evaluation data that may have potential utility for assessing change in E&A programmes.

What needs considering during implementation?

Key messages:

- Plan an inception workshop to iron out the details of the MEL operational plan
- Develop protocols with adequate capacity development for those involved
- In addition to planning a final evaluation, ensure resources are available for periodically revisiting TOCs, interrogating assumptions, analysing contributions with reference to context, and adapting programmes in light of learning.

[5] www.bond.org.uk/data/files/Effectiveness_Programme/IIF_thematic_papers/Empowerment.pdf
www.bond.org.uk/data/files/Effectiveness_Programme/IIF_thematic_papers/GovernanceAccountability.pdf
(both accessed 21 March 2013)

[6] http://xa.yimg.com/kq/groups/18751725/128169439/name/1LikertScales.pdf (accessed 14 March 2013)

In many instances, the contexts in which programmes are implemented and the precise activities planned will change between programme design and the beginning of implementation. Hence MEL strategies will need revisiting, refining and developing into more detailed operational plans. This will require a framework that clearly describes who is going to do what and when to collect data, analyse it and apply learning to change. If integrated in a participatory design, the plan might be an output of a workshop.

Key questions for a MEL inception workshop

- What does the context analysis suggest about the validity of our assumptions and initial theory of change?
- What are our final definitions of outcomes and impacts? What are the indicators; data sources; and tools we will use to assess change?
- When and where are we going to collect data? What are the likely opportunity costs for women and marginalised groups? Will they be able to participate?
- Who is responsible for collecting and storing, processing data, including any generated through meetings, scorecard processes or administrative records?
- Who is responsible for 'quality control', technical consolidation and analysis?
- Who will participate in interrogating the TOC, when will this happen?
- How will we ensure what we learn results in decisions that are applied to programme management and activities?

All MEL approaches need to develop protocols and guidelines for data collection to ensure tools are used and data is collected in a fairly standard way. Training of facilitators needs time and financial resources; budgets must be considered early on.

Example of a checklist for fieldwork

- Is there a simple field manual that specifies the tools to be used and the steps to be followed including those related to ethical issues?
- Do the research teams have adequate facilitation, interviewing and note-taking skills; enthusiasm; the language skills? Are they of an age and gender conducive to establishing a good rapport with research participants?
- If participants/respondents in participatory research are selected by voluntary rather than random sampling approaches, is there a protocol for exploring who attends, why and what this means in terms of representation? Do they stress the need to seek out unrepresented groups for informal interviews?
- How will sensitive issues that could lead to a response bias in groups – people echoing the views of the most powerful – be handled?
- If working with small 'n' samples or case studies, do we have protocols to collect names of individuals so we can revisit them to explore longitudinal change?
- Do we have note-taking protocols for assessing the quality of research encounters e.g. level of participation during the discussion?
- Do we have protocols to increase the credibility of data during field work, e.g.:

    - *Ad hoc* efforts to triangulate through informal discussions with reliable key informants or people from special groups
    - Immediate debriefing – facilitators share results from focus groups with peers who know the situation and have to defend the data
    - Local-level analysis and interpretation by those familiar with the context
    - Sharing back consolidated results for empowerment and validation.

What needs to be considered when designing analysis approaches? Data must be trusted. 'Real-time' action research approaches to M&E that encourage participatory analysis and critical reflection with relation to TOCs that are constantly updated throughout the project are likely to enable more robust analysis of data generated during key mid-term and final evaluation fieldwork.

The nature of analysis carried out on the data at different times in the programme depends on the type of data and the objectives of the study, not on the tools used to collect the information. Coding, analysis and sense-making will be influenced by discussions about theories of change etc. In complex programmes data analysis should be iterative. The first stage is exploratory and involves the use of simple descriptive statistics for quantitative data and codes for the analysis of qualitative data. The second stage involves more detailed interrogation of what findings reveal about the validity of theories of change and the relative contributions of different factors and actors to change or lack of it using a process such as that described in the box below.

Contribution analysis and testing theories of change

Contribution analysis has traditionally been described as exploring attribution through assessing the contribution a programme is making to observed outcomes. It sets out to explore M&E data to verify/test the theory of change behind a programme and, at the same time, takes into consideration other influencing factors. Programmes using real-time adaptive management processes can undertake contribution analysis for different components of programmes at different levels and at different points in time, e.g. after a period when it is estimated there will be some outcomes emerging. In some instances it may be appropriate to focus on in-depth case studies in a few carefully and purposively chosen sites of interest, as a means to generate real-time learning that can be used to pilot test and revise the TOC and broader approach across a larger area.

To apply contribution analysis to a theory-based E&A evaluation take the following steps:

- Revisit the theory of change and assumptions specified, using a participatory process if using a hybrid participatory design. Remind everyone of the assumed causal mechanisms being explored.
- Collate original and current context and power analysis data, baseline data, monitoring data, quantitative and qualitative evaluation data.
- Map the quantitative and qualitative data onto a theory of change schema using a chronological sequencing to help explore 'causal' mechanisms. Try and triangulate different data, or data from different sources if and where possible.
- Test whether consolidated data validate the TOC and assumptions.
- Use context analysis to explore other factors that might have influenced change and alternative explanations from those proposed in original assumptions. Examine alternative explanations seeking to *disprove* – not to prove – expectations that 'your' actor contributed significantly to the change.
- Assemble and assess the contribution story and challenges to it. Seek additional evidence if required and revise and strengthen the contribution story.
- Feedback consolidated data to broader stakeholder groups to empower, mobilise, and facilitate mutual responsibility between various actors.
- Ask questions: do our assumptions hold? Given new understandings of context and emerging results, are we doing the right thing?
- Adapt programmes as a result of learning.

(Adapted from www.cgiar-ilac.org/files/publications/briefs/ILAC_Brief16_Contribution_Analysis.pdf)

What to consider when writing or reviewing reports?

Key messages:

- Make evident how, when and why evaluation decisions were taken, e.g. design, sample size, approach to analysis
- Discuss limitations, doubts, and the positionality or bias of the researchers
- Ensure conclusions, especially those that relate to validity of assumptions and TOCs discuss the implications for programmes elsewhere.

Even with the best planning, MEL in E&A programmes is always going to be complex and challenging. All methodologies will have their own sets of limitations and things will undoubtedly go wrong. NGOs' increasing interest in improving the quality of evidence produced by their evaluations means there is an urgent need to contract people with – or build internally – the confidence and ability to write reports that include methodology sections that are less descriptive, more analytical, and critically reflexive about limitations. Although coming from a research paradigm not usually applied to E&A programmes, the methodological discussion in Care's randomised control trial in the Democratic Republic of Congo (Humphreys *et al.* 2012) is an example; so too is that included in the Care Bangladesh Participatory Impact Assessment report (Gillingham 2011).

Clear structuring and detailed discussions of findings, analysis and conclusions in reports greatly enhances readers' perceptions of reliability and credibility and their ability to learn from the experiences of others. Theory-based evaluation approaches require integrating quantitative and qualitative data in a change narrative that carefully unpacks and discusses the context, causal factors and change mechanisms. The World Vision Armenia report (2011) and Christian Aid GTF Mid-term Review (McGee and Scott-Villiers 2011) were some of the only evaluations reviewed that started to explore the influence of context on outcomes. This is not surprising, as it can be difficult to achieve in standard evaluation reporting formats that only require sections on efficiency, effectiveness etc. NGOs need to take advantage of skills they have developed writing anecdotal case studies and apply the same approaches to writing contextually nuanced cases or stories that include more 'robust' evidence collected through some of the processes above. These can be included as appendices (cross-referenced in the main text) if reporting formats make it difficult to present findings in ways that communicate stories of change or lack of it.

Tool 4: Glossary of terms used in MEL debates and documents

Causal inference	Conclusion that a cause is linked to an effect
Counterfactual	Comparison of what actually happened with what would have happened without an intervention
Descriptive statistics	Describe in quantitative terms the main features of data (as opposed to inferential statistics that try to support more general conclusions using statistical tests)
Evaluation design	Overarching logic for evaluations. Includes: questions, theory used to analyse data, data and use of data
Experimental design	Evaluation design developed in the natural and medical sciences. A 'treatment' or intervention is applied to a subject or group of subjects, and observations made of what happens to subject(s) are compared (through statistical analysis) with observations of a 'control group' or counterfactual that is isolated from the intervention, e.g. as in a randomised control trial, (RCT). In principle, because the 'subjects' of the treatment are selected randomly, the only difference between them and the control group is that the intervention has been applied to the former.

Likert scale	Rating tool use to scale responses in surveys – eg 'Rank the following from 1–5 according to how highly you prioritise them'
'Most Significant Change'	A story-based participatory technique used to help improve programmes by including participants in data collection and analysis to enable learning to focus the direction of work towards directions explicitly valued by participants
Quasi-experimental design	Similar to experimental designs but lacking random assignment to treatment or control groups. The researcher uses different criteria to select a suitable group or situation for comparison
Realist evaluation	Variant of theory-based evaluation used in areas where no established programme theory exists. Used to build theories
Real-time evaluation	An evaluation that is conducted during an intervention in order to learn and adapt to enhance impact
Small 'n' study	Study that includes a small sample size, sometimes used to describe case study approaches ('n' is the statistical term for the number of cases under analysis)
Theory of change	Programme planning and evaluation tool used to describe causal pathways and assumptions about how change happens in projects and programmes
Theory-based evaluation	An approach to evaluation that tests theories underpinning programmes, e.g. theories of change or programme theories
Triangulation	Use of two or more methods or data sources to validate the same findings or results

Bibliography

Adams, J. (2012) *Using Scalar Approaches to Monitor Advocacy and Empowerment Work: Best Practice Paper*, PPA Learning Group on Measuring Results in Empowerment and Accountability, INTRAC

Argyris, C., and Schön, D. (1978) *Organizational Learning: A Theory of Action Perspective*, Reading MA: Addison Wesley

Barahona, C. and Levy, S. (2002) *How to Generate Statistics and Influence Policy using Participatory Methods in Research,* Statistical Services Centre Working Paper, University of Reading

Barder O. (2012) *Global Development: Views from the Centre – Complexity, Development and Results,* http://blogs.cgdev.org/globaldevelopment/2012/09/complexity-and-results.php (accessed 20 October 2012)

Barder, O. and Ramalingam, B. (2012) *Complexity, Adaptation and Results*, blogpost, http://blogs.cgdev.org/globaldevelopment/2012/09/complexity-and-results.php (accessed 1 November 2012)

Chambers, R. (1987) *Rural Development: Putting the Last First*, Longman UK

Davies, R. and Dart, J. (2005) 'The Most Significant Change Technique: A Guide to its Use', in CARE *et al., ActionAid International Critical Stories of Change*, www.actionaid.org/main.aspx?PageID=894 (accessed 31 March 2009)

Jamieson, S. (2004) 'Likert Scales How to (Ab)use Them', *Medical Education* 38: 1212–8 http://xa.yimg.com/kq/groups/18751725/128169439/name/1LikertScales.pdf (accessed 14 March 2013)

Gillingham, S. (2011) *Understanding Change: A Participatory Impact Assessment of the SETU Project's Community-Led Development Approach*, CARE

Holland, J. (2010) *Good Practice Note: Monitoring and Evaluating Empowerment Processes*, Berne: Swiss Development Cooperation

Hughes, K. and Hutchings, C. (2011) *Can we Obtain the Required Rigour without Randomisation? Oxfam GB's Non-experimental Global Performance Framework*, Working Paper 13, International Initiative for Impact Evaluation

Humphreys, M.; Sanchez de la Sierra, R. and van der Windt, P. (2012) *Social and Economic Impacts of Tuungane,* Final Report on the Effects of a Community Driven Reconstruction Program in Eastern Democratic Republic of Congo, New York: Columbia University

Jupp, D. and Ibn Ali, S., with Barahona, C. (2010) 'Measuring Empowerment? Ask Them – Quantifying Qualitative Outcomes from People's Own Analysis', *Sida Evaluation Series* 1, Stockholm: Sida, www.oecd.org/countries/bangladesh/46146440.pdf (accessed 21 March 2013)

Roche, C. and Kelly, L. (2012) *The Evaluation of Politics and the Politics of Evaluation*, DLP Background Paper 11, Developmental Leadership Program, summary available at www.gsdrc.org/go/display&type =Document&id=4340 (accessed 21 March 2013)

Mayne, J. (n.d.) *Contribution Analysis: An Approach to Exploring Cause and Effect*, www.cgiar-ilac.org/files/publications/briefs/ILAC_Brief16_Contribution_Analysis.pdf (accessed 15 March 2013)

McGee, R. and Gaventa, J. (2011) *Shifting Power: Assessing the Impact of Transparency and Accountability Initiatives*, IDS Working Paper 383, Brighton: IDS

McGee, R. and Gaventa, J. (2010) *Review of Impact and Effectiveness of Transparency and Accountability Initiatives*, Brighton: IDS

McGee, R. and Scott-Villiers, P. (2011) *GTF Mid-term Review*, Christian Aid

Pawson, R. (2002) 'Realist Synthesis: Supplementary Reading 2 – "In Search of a Method" ', ESRC Centre for Evidence-Based Policy, Queen Mary College, University of London. *Evaluation*, 8.2: 157–81, http://evi.sagepub.com/content/8/2/157.abstract (accessed 1 November 2012)

Ramalingam, B. and Jones, H. with Reba, T. and Young, J. (2008) *Exploring the Science of Complexity: Ideas and Implications for Development and Humanitarian Efforts*, ODI Working Paper 285, London: Overseas Development Institute

Reeler, D. (2007) *A Threefold Theory of Social Change and Implications for Practice, Planning, Monitoring and Evaluation*, www.cdra.org.za/articles/A%20Theory%20of%20Social%20Change%20by%20Doug%20Reeler.pdf (accessed 21 November 2012)

Roche, C. and Kelly, L. (2012) *The Evaluation of Politics and the Politics of Evaluation*, DLP Background Paper 11, Developmental Leadership Program (summary available at www.gsdrc.org/go/display&type=Document&id=4340 – accessed 15 March 2013)

Rondinelli, D. (1993) *Development Projects as Policy Experiments: An Adaptive Approach to Development Administration*, Routledge

Shutt, C. (2010) *Monitoring, Evaluating and Assessing the Impact of Governance Programmes: Summary of Literature and Practice Review* unpublished report from available from Care UK

SSC (2001) *Some Basic Ideas of Sampling, Statistical Good Practice Guidelines*, Reading: University of Reading, Statistical Services Centre

Stern, E.; Stame, N.; Mayne, J.; Forss, K.; Davies, R. and Befani, B. (2012) *Broadening the Range of Designs and Methods for Impact Evaluations*, DFID Working Paper 38, London: DFID

Waswaga, R.; Winterford, K.; Walker, B.; Wamala Mugabi, B. and Otim, B. (2011) *Citizen Voice and Action End of Pilot Project Evaluation*, unpublished report available from World Vision

World Vision (2011) *End of Project Evaluation Report: Policy Influence through Community Empowerment*, World Vision Armenia

York, N. and Hoy, C. (2012) *What do DFID Wonks think of Oxfam's Attempt to Measure its Effectiveness* www.oxfamblogs.org/fp2p/?p=12254 (accessed 15 March 2013)

This CDI Practice Paper Annex was written by **Cathy Shutt** and **Rosie McGee**.

Acknowledgements The authors would like to thank Jennifer Doherty (Plan UK), Nicole Walshe (CARE UK), Gaia Gozzo (CARE UK), Jake Allen (Christian Aid), and Daniel Stevens (World Vision UK) for helping shape this output. We are also grateful to Jake Phelan (Plan UK), Hilary Williams (World Vision UK), John Lakeman (CARE UK), Francesca Scott (Christian Aid), Ariel Frisancho Arroyo (CARE Peru), Donald Mogeni (World Vision UK) and Bill Walker (World Vision Australia) for sharing their time, experience and ideas that improved the report.