METHODS
LAB

March 2016

EVALUATING THE IMPACT OF FLEXIBLE DEVELOPMENT INTERVENTIONS USING A 'LOOSE' THEORY OF CHANGE

REFLECTIONS ON THE AUSTRALIA-MEKONG NGO ENGAGEMENT PLATFORM

Dr Rick Davies

KEY MESSAGES

- For some interventions, tight and testable theories of change are not appropriate – for example, in fast moving humanitarian emergencies or participatory development programmes, a more flexible approach is needed.
- However, it is still possible to have a flexible project design and to draw conclusions about causal attribution. This middle path involves 'loose' theories of change, where activities and outcomes may be known, but the likely causal links between them are not yet clear.
- In this approach, data is collected 'after the event' and analysed across and within cases, developing testable models for 'what works'. More data will likely be needed than for projects with a 'tight' theory of change, as there is a wider range of relationships between interventions and outcomes to analyse. The theory of change plays an important role in guiding the selection of data types.
- While loose theories of change are useful to identify long term impacts, this approach can also support short cycle learning about the effectiveness of specific activities being implemented within a project's lifespan.

Dr Rick Davies is an independent monitoring and evaluation consultant.

This paper was reviewed by Irene Guijt (Overseas Development Institute) and Sarah Ransom (Australian Embassy Hanoi).

A METHODS LAB PUBLICATION

ODI.ORG/METHODSLAB

The Methods Lab is an action-learning collaboration between the Overseas Development Institute (ODI), BetterEvaluation (BE) and the Australian Department of Foreign Affairs and Trade (DFAT). The Methods Lab seeks to develop, test, and institutionalise flexible approaches to impact evaluations. It focuses on interventions which are harder to evaluate because of their diversity and complexity or where traditional impact evaluation approaches may not be feasible or appropriate, with the broader aim of identifying lessons with wider application potential.

The views presented in this paper are those of the author(s) and do not necessarily represent the views of ODI, the Australian Department of Foreign Affairs and Trade (DFAT) and BetterEvaluation.

How to cite this working paper:
Davies, R. (2016) ‘Evaluating the impact of flexible development interventions using a ‘loose’ theory of change’. A Methods Lab publication. London: Overseas Development Institute.

Research
& Policy in
Development

BetterEvaluation

Australian Government
Department of Foreign Affairs and Trade

Overseas Development Institute
203 Blackfriars Road
London SE1 8NJ
Tel +44 (0) 20 7922 0300
Fax +44 (0) 20 7922 0399
info@odi.org.uk
www.odi.org

BetterEvaluation
E-mail: bettereval@gmail.com
www.betterevaluation.org

Contents

Acknowledgements

The author thanks Sarah Ransom for her consent and support in writing about the AMNEP programme; Irene Guijt for her constructively critical review of drafts of this paper; Tiina Pasanen and Hannah Caddick for their careful and through editorial assistance.

Acronyms

AMIS	Automated management information system
AMNEP	Australia-Mekong NGO Engagement Platform
DAC	Development Assistance Committee
DFAT	Department of Foreign Affairs and Trade, Australia
DFID	Department for International Development, UK
M&E	Monitoring and evaluation
MEL	Monitoring, evaluation and learning
NGOs	Non-governmental organisations
OECD	Organisation for Economic-Cooperation and Development
QAI	Quality at Implementation
QCA	Qualitative Comparative Analysis
RCTs	Randomised control trials

1. A big ask: flexible design with causal attribution

This paper looks at the challenge of evaluating the impact of projects that aim to be flexible and responsive, with reference to one such project, funded by DFAT,[1] the Australia-Mekong NGO Engagement Platform (AMNEP). A core characteristic of an impact evaluation is the effort to establish attribution, i.e. the causal role of interventions designed to have an impact on people's lives or the institutions that affect them.[2] One of the criteria for good impact evaluation is rigour – which, broadly translated, means having a transparent, defensible and replicable process of data collection and analysis. And its debatable apotheosis is the use of randomised control trials (RCTs). Using RCTs requires careful management throughout the planning, implementation and evaluation cycle of a development intervention. However, these requirements for control are the antithesis of what is needed for responsive and adaptive programming. Less demanding and more common alternatives to RCTs are theory-led evaluations using mixed methods. But these can also be problematic because ideally a good theory contains testable hypotheses about what will happen, which are defined in advance.

The push for more adaptive programming comes from multiple directions. Partly from a continuing tradition within development that emphasises the importance of people's participation, most recently represented within the Big Push Forward initiative.[3] Partly from pragmatists within aid agencies (e.g. DFID, World Bank) that have seen at first hand the limitations of blueprint based projects (Vowles 2015). And partly from a steady stream of writers on the subject of complexity and systems thinking, who emphasise the unpredictability of events (Ramalingam 2013, Snowden 2015). The latter are, almost by definition, not very optimistic about the value of clearly articulated theory of change to generate accurate predictions.

The approach encouraged in this paper is to find a middle way between relying on pre-defined theories of change and abandoning any hope altogether that they can cope with the open-ended nature of development. Practical ways of doing this can be borrowed and adapted from the commercial sector – specifically a growing body of knowledge and practice known as predictive analytics.[4] In this sphere, predictive models of consumers' behaviour (for example) are developed *after the event*, but based on extensive data sets of what is known that did happen.

This paper argues that the value of this data-centred approach can be enriched at two key stages by two more theory-oriented approaches, adapted from Qualitative Comparative Analysis (QCA) (Ragin 1989). First, when making decisions about what kinds of data to collect, there is a useful role for 'loose' theories of change of the kind described in this paper. The second stage is the point at which associations between events and outcomes are elaborated into plausible causal claims through the subsequent use of more qualitative within-case inquiries into key cases.[5]

In reality, implementing such an approach in real-life circumstances is not necessarily straightforward. This paper reviews the short history of AMNEP through the eyes of a monitoring and evaluation (M&E) consultant who thought that, in this case, the development of after-the-event models of what did happen would be most appropriate approach to ensuring both learning and accountability through impact evaluation.

After providing some background to AMNEP this paper looks at challenges in four areas:

- the definition of objectives
- the measurement of outcomes
- data storage
- data analysis

We focus primarily on the early stages of the data collection and analysis process described above. Due to the premature termination of AMNEP in mid-2015 we were unable to complete the data collection and analysis stage or to make follow-up, within-case inquiries. But we are able to conclude with some suggestions of 'lessons to be learned'.

[1] The Department for Foreign Affairs and Trade, of the Government of Australia.

[2] 'Impact evaluation is an assessment of how the intervention being evaluated affects outcomes, whether these effects are intended or unintended' www.oecd.org/dac/evaluation/dcdndep/37671602.pdf.

[3] www.ids.ac.uk/project/the-big-push-forward.

[4] Siegel, E. (2014) Predictive Analytics. John Wiley and Sons.

[5] Cases is a generic term for units of analysis, which might for example in AMNEP refer to different projects being assisted.

2. The context: the Australia-Mekong NGO Engagement Platform

The AMNEP aimed to improve the effectiveness of Australian aid through non-governmental organisations (NGOs) and improve partnerships between DFAT and NGOs.[6] It sought to achieve these outcomes through two 'tracks' of activities: (a) the provision of technical assistance to bilateral and regional projects being implemented by NGOs; and (b) the provision of new opportunities for policy dialogue and learning between DFAT and NGOs. The underlying strategic assumption was that this support would improve the relevance and quality of NGO activities, and thus lead to better impacts for citizens in the Mekong. The AMNEP did not function as a funding body; instead it was a provider of optional services to DFAT staff administering various DFAT funds and to NGOs implementing development projects using those funds.

The AMNEP was approved for funding in August 2012, with a budget of AU$15 million, for an initial six-year period. The AMNEP began interim operations from Canberra in late 2012 and moved to full implementation in June 2013. In September 2013 a decision was made to merge AusAID into DFAT. Subsequently the Australian aid budget was cut back in 2014 and 2015, and as part of that process, a decision was made in mid-2015 to bring the AMNEP to a close by the end of 2015.

[6] This paragraph's description is the most current, being based on text in the AMNEP's 2015 QAI report.

3. Putting a loose theory approach into practice

3.1. Clarifying the purpose of evaluation: accountability and/or learning

The Monitoring, Evaluation and Learning (MEL) Framework included in the AMNEP's Final Design Document (2012) did not include any specific targets to be achieved at the output, outcome or impact level; the text descriptions simply referred to events (plural) of various types. This means it was not possible for the AMNEP to be held to account in any simple target-oriented way. The AMNEP did produce an annual work plan but this was inherently limited in its coverage given that 50 per cent of resources were intended to be available on a responsive, rather than scheduled, basis. Detailed monthly update reports were produced along with a six-monthly financial report, but these were focused on the AMNEP outputs rather than outcomes and impacts. A wider view on achievements was required through the completion of annual Quality at Implementation (QAI) reports.

The 2013 draft M&E framework responded to the lack of specific targets by recommending that the nature of the AMNEP's long-term accountability be cast in terms of responsibility to generate knowledge about what works in what circumstances. That is, what types of AMNEP technical assistance and dialogue promoting activities have a positive impact on aid effectiveness and DFAT-NGO relationships, and under what conditions? Given these objectives it was appropriate to consider a form of impact evaluation that aimed to uncover causal relationships.

3.2. Assessing evaluability

Evaluability has been defined by the Organisation for Economic Co-operation and Development's Development Assistance Committee (OECD-DAC) as 'The extent to which an activity or project can be evaluated in a reliable and credible fashion' (OECD-DAC 2010, p.21). Checking if a development intervention is evaluable (Davies, 2013; Peersman, et al. 2015) prior to initiating an impact evaluation should be common sense: failing to check if an impact evaluation is, in fact, feasible can lead to a poorly timed evaluation, invalid findings and lack of utility. In short, it is a waste of time and resources.

The author's initial assessment of the AMNEP's design in mid-2013 was that it was unevaluable. This view was based on a reading of the Final Design Document and a diagrammatic version of the theory of change, developed shortly after the AMNEP's approval, which revealed two major problems. The first, present in the Final Design Document, was the lack of clarity about the expected temporal sequence of events that would take place as the AMNEP was implemented. Unexpectedly, the purpose-level descriptions included references to 'AMNEP flexibly responding to sector program opportunities,' something that would normally be expected much earlier in a causal pathway described by a theory of change. Meanwhile, what would conventionally be described as project outputs that fed into the purpose were absent and in their place was a set of 'domains of change', built around OECD-DAC evaluation criteria.

The second problem was most evident in the diagrammatic version of the theory of change subsequently developed in 2013, shortly after AMNEP's approval. This was the lack of identifiable (and therefore testable) causal links between the 50 or more events taking place at what now appear to be five different levels. This assessment of un-evaluability was shared by the recently appointed AMNEP coordinator, who had also expressed more immediate concerns about the breadth of ambition and the implementability of the design.

This situation should be seen in context. The approval and funding of projects that are initially unevaluable is not uncommon in aid organisations generally. Development projects may be approved for many reasons, often due to the politics of aid as much as (or more than) their coherence as a technical proposition. Or, they may be approved due to their recognition of their novelty and therefore a willingness to accept a period of refinement.

Inception periods are often about clarification and consensus-building on project purpose as well as what interventions might feasibly achieve it. This is especially the case with larger projects, which have a diverse set of stakeholders. Prior to the merging in 2013 of AusAID with DFAT, evaluability assessments were used on an almost routine basis by the AusAID Indonesia Aid Program in the early inception stages of the project cycle to further articulate project designs. Evaluability assessments have also become more widely used by other development agencies in the last ten years or so, in de facto recognition of the fact that the official approval of a project is not the actual end point in project design (Davies 2013). In the case of the AMNEP, the use of an evaluability assessment to feed into the design of an M&E framework was a practical alternative following a prolonged design phase and evidence of 'design fatigue'.

If post-approval un-evaluability is common then how should such problems be remedied once identified? In the

AMNEP case it is likely that at least some of the problems could have been avoided if a conventional logframe structure had been used, as had been in AusAID projects in the past. A temporal sequence of events might have been spelled out and some bridging assumptions attended to. Ideally this would have distinguished between events that the AMNEP was responsible for and their effects, i.e. the difference between activities and outputs versus outcomes and impacts.[7]

However, using a logframe structure is only a partial solution; it provides some overall order but leaves much uncertainty. Within each level of most logframes – especially at output and outcome levels – there are typically multiple events whose temporal sequencing and connecting causal linkages are left to be detailed by one means or another. In some contexts these might be resolved during an inception period and then assessed during an evaluability assessment (Davies 2013, Peersman et al. 2015). However, in projects like the AMNEP, uncertainty (i.e. openness) about what AMNEP support would be provided, to whom and when, was an integral part of its design and not a gap in it. This 'looseness' presents a major challenge to a theory-led evaluation approach, which ideally requires hypotheses to be stated up front, and data gathered and then used to test them.

This becomes less of a challenge if the focus is on retrospective analysis of the relationships between these events after they have happened. But to carry out such an analysis on a systematic and comprehensive basis requires much more data. This is because data needs to be available on all the events within a loose theory of change that could be making some form of causal contribution, and not only a sub-set associated with a specific prior hypothesis about how change will work. If this kind of ex-post analysis is possible, then what was an unevaluable project may now be evaluable.

This is a significant development, substantially widening the scope of impact evaluations. The key point here, and discussed in detail later in this paper, is the possibility of *systematically* exploring all possible combinations of events in a loose theory of change. This is unlike the selective testing of favoured hypotheses identified once data has been collected, a practice which is vulnerable to accusations of 'cherry picking', involving the 'fallacy of selective attention'.[8]

3.3. Developing clear and realistic expectations about expected impacts

AMNEP's Final Design Document summarised the AMNEP's highest level objective ('goal') as being to 'Achieve a better quality aid program in the Mekong'. Its achievement would be evidenced by: (a) 'Strengthened AusAID policies and programs through AusAID – NGO engagement'; and (b) 'Uptake of policy and change in practice by other actors (for example, partner governments) based on learning and experience of AMNEP partners'. The AMNEP's intermediate objective ('purpose') was to 'Change in the way in which AusAID and NGOs do business'. The project clearly had a broad ambit, covering projects implemented by NGOs and those directly managed by AusAID. And it also aimed to influence partner governments as well as affect the way AusAID and NGOs worked together.

The scale of this ambition presented two impact evaluation challenges. One was the wide range of projects and associated stakeholders where the AMNEP's impact would need to be identified – in 2012-13 DFAT spent AU$327 million on 59 different projects spread across four countries.[9] The other challenge was the likely difficulty of identifying impacts on aid effectiveness given the small scale of AMNEP inputs into these projects – its annual budget was equivalent to 0.8 per cent of total DFAT expenditure in the region.

During its first two years' implementation, the AMNEP's statement of objectives was progressively refined and became more realistic. The first objective became more clearly focused on improving the aid effectiveness of DFAT-funded NGO projects;[10] by the end of 2015 the AMNEP had focused in on a much smaller sub-set of 14 projects. Under the second objective the role of other actors (e.g. partner governments) receded and the project's objective statements concentrated on working relationships between DFAT and NGOs. Both changes made the task of monitoring and evaluating impact less challenging because the range of actors was much reduced. There was, however, still considerable 'looseness' in the project's design with respect to which DFAT-NGO working relationships would be of primary concern. In practice, these were selected on the basis of shared interests in different development topics, which were the focus of a series of regional forums.

[7] The same distinction is made in outcome mapping terminology, between sphere of control versus sphere of influence and interest.

[8] https://en.wikipedia.org/wiki/Cherry_picking_(fallacy)

[9] This leaves out Australian NGO projects in these countries funded by DFAT through Australia-based funding channels.

[10] Including bilateral projects involving NGO as partners.

During the three years over which the project was implemented, the proposals made for M&E were responsive, being led by changes in the perceived objectives of the project that were driven by other forces. Contrary to some concerns expressed about development projects elsewhere (Eyben et al. 2015) it was not the case with the AMNEP that design and implementation was driven by what was measurable and demands for measurable results being dictated by funding sources.

The narrowing and refining of the AMNEP objectives was a step in the right direction. The data collection requirements implied by the original objectives had been reduced dramatically and those implied by the most recent objective statements were more within reach. The next section looks at the kinds of data that were seen as being both practically accessible and useful.

3.4 Measuring impact-level changes

Aid effectiveness

Ideally the impact of improved aid effectiveness would be visible in the form of improvements in the lives of those people intended to benefit from Australian aid in the Mekong region. In reality, the task of measuring these changes (let alone making any claims about causal attribution) was fraught with difficulty. Both across and within each of the finally selected 14 projects, the range of types of intended beneficiaries and the expected changes in their lives varied widely. Because such impact on would be a temporally distant event relative to the timing of the AMNEP assistance it was likely that there would be countless other forces contributing to how those projects finally affected people's lives. Comparability of outcomes, and the ability to make any claims about what caused these, will be very limited. And if the main source of information about impact on people's lives was likely to be project evaluations these would be few in number and undertaken at different points in time.

One way around this problem was to look for more proximate and proxy measures of impact. This type was available in the form of an annual assessment of AusAID projects, known as Quality at Implementation (QAI) reviews. These were produced for all AusAID projects with a value of more than AU$3 million or deemed otherwise 'significant', and have been produced since 2008. The presentation and procedure for completion of QAI reviews followed a standard format, which included the use of internal peer review and an annual 'spot-check' review carried out by external consultants.[11]

The QAI reviews covered multiple dimensions of performance: relevance, efficiency, effectiveness, M&E, sustainability and gender equality. Each dimension was rated on a six-point scale with the requirement for a supporting narrative argument and evidence. QAIs were undertaken annually and were followed by a Quality at Completion assessment, which forms part of the Independent Completion Report. In summary, QAIs offered three advantages: a standardised measurement instrument, covering multiple aspects of performance, covering multiple points in time in the lifecycle of a project.

But concerns about this approach were expressed by DFAT staff. One was the *insensitivity* of the instruments. Although QAI judgements used a six-point scale, in one sample of QAI-reviewed DFAT projects, 80 per cent of the ratings of effectiveness used only two of the points in the scale. In other words, there were few differences between the projects so the possibility of distinguishing between those assisted by the AMNEP and those not appeared very limited. However, when ratings on all six quality scales were aggregated for each project, there was much wider range variation in overall performance of the projects. The QAI then looked like it could be sensitive enough. It was through the use of this looser conception of 'quality' – in the form of an aggregation of the six different scales – which provided improved sensitivity.

Another concern was the *availability* of the QAI data itself, especially once AusAID had been absorbed into DFAT in late 2014. Fortunately, the production of country-specific Aid Program Performance Reports continued during this transition. These reports include QAI rating data on all projects on all six criteria for all projects over AU$3 million in value.

Consistency of data emerged as an unexpected issue. After the incorporation of AusAID into DFAT, the QAI instrument was revised – primarily with the aim of reducing the amount of staff time needed – which could have had implications on the consistency of the data. Somewhat surprisingly comparisons of ratings on specific criteria for specific projects, before and after this change have not revealed any noticeable change in the balance of judgements.

This may be because of the compensating benefits of another aspect of the revision. For each of the six aid-quality criteria that still had to be assessed there was now a set of standardised subsidiary questions, each with its own six-point scale for answers (see the example in Figure 1). This finer 'granularity' has increased the potential for identifying differences in performance results between DFAT projects, even on individual QAI criteria. It is important to note that there is no one correct way in which the scores on the subsidiary scales

[11] Review and Spot Check of Quality at Implementation (QAI) Reports. Office of Development Effectiveness. AusAID 2013.

should be aggregated: how this is done has been left to the judgement of the user, which can then be explained in an associated comment field.

Because of this freedom, each of the six aid-quality criteria now have their own ‘loose’ theory, i.e. one where the contribution of each subsidiary question and scale to the overall rating for that aid criteria was not prescribed in advance but by each user of the QAI. This innovation in effect provides another opportunity to learn from after-the-event analysis of data about ‘what works’, this time at a more micro level. Analysis could shed light on which of the subsidiary questions (and combinations thereof) had the most influence over aggregate judgements that respondents made on the QAI criteria they relate to.

Partnership, dialogue and mutual learning

As noted, the second objective of the AMNEP was refined to refer to improved relationships between DFAT and NGOs. The difficulty of measuring the type of change to which the AMNEP wants to contribute has been compounded by the fact that the terms used to describe this objective have changed over time. In the Final Design Document there were many references to *partnership*, whereas since the merger with DFAT *dialogue* and *mutual learning* have been the preferred terms in written documents. In the Final Design Document the concept of partnership is described in ambivalent terms, both normatively as a particular kind of desired relationship and more neutrally as an array of different kinds of working relationships.

The AMNEP Coordinator’s first interpretation of partnership took the form of a four-part typology, which also formed a scale, ranging from information generation and sharing; to consultation; to collaboration; to partnership. Within each type there were degrees of intensity, observable during both design and implementation stages of a project cycle, and which were identifiable with the aid of a rubric. There are both advantages and disadvantages to the use of this kind of structured and annotated scale. On the one hand they can help make initially nebulous concepts more observable and measurable. On the other, they bind the user into a relatively tight theory, i.e. that working relationships can be located somewhere on this linear scale and that this location will then have some correlation with performance in other areas, such as aid effectiveness.

Subsequently, in late 2013, the AMNEP Coordinator revisited the ways in which partnership could be assessed by securing a brief literature review on the subject from the UK government Department for International Development (DFID)-funded Governance and Social Development Resource Centre. They reported: ‘There are not many specific tools available, as most organisations rely on generic internal feedback and consultation sessions, rather than comprehensive monitoring and evaluation of relationships.’ Six tools, used by agencies such as Keystone, Bond and World

Figure 1: A ‘loose’ theory of what makes a project ‘effective’

Effectiveness: are we achieving the results that we expected at this point in time?						
Aid quality criteria for assessment	**Six-point scale**					
The investment has clear and realistic outcomes, supported by a robust logic and theory of change	1	2	3	4	5	6
The investment is on-track towards achieving its expected outcomes	1	2	3	4	5	6
The quality of the investment’s key outputs and activities is as expected	1	2	3	4	5	6
Policy dialogue is used effectively to influence partners and support the investment’s outcomes	1	2	3	4	5	6
Intended beneficiaries are satisfied with the investment’s results	1	2	3	4	5	6
The investment actively involves disables peoples’ organisations in planning, implementation and monitoring and evaluation	1	2	3	4	5	6
The investment identifies and addresses barriers to inclusion and opportunities for participation for people with disability	1	2	3	4	5	6
The investment identifies and addresses barriers to inclusion and opportunities for participation by indigenous peoples or ethnic minorities	1	2	3	4	5	6
Overall rating	**1**	**2**	**3**	**4**	**5**	**6**

Wildlife Fund, were described in some detail but none could be directly applied in the AMNEP context without adaptation.

The following year these tools were used by the M&E consultant as the basis for a composite relationship survey instrument tailored to the needs of the AMNEP. This instrument sought respondents' judgements on what are called 'facet' and 'global' judgements of the respondents' working relationships between DFAT and NGOs.[12] 'Facets' are specific aspects of a relationship, such as trust or confidence; 'global' judgements are those that refer to 'overall satisfaction' or similar. However, the instrument made no prior assumptions (e.g. via the use of weightings) about how much each facet would contribute to the respondents' global (i.e. overall) judgements about their satisfaction with the relationship.

This mixture of a set of facet measures and one global measure embodied a 'loose' theory about what mattered in the working relationships between DFAT and NGOs. Plausible casual connections between the facet and global aspects of the surveyed relationships could only be identified after data had been collected from respondents. An important difference between the Coordinator's initial measurement scale and this loose theory is that the latter is a multi-dimensional combinatorial space in which many more types of partnerships can be found, and any one (or more) of these may be associated with positive development outcomes.[13]

3.5. Measuring short-term outcomes

There were two short-term outcomes that were expected to be causally related to the longer-term outcomes already discussed:

1. the effects of the AMNEP technical assistance, as perceived by the immediate clients of that assistance
2. the value of the AMNEP-facilitated policy dialogues and learning events, as perceived by the participants.

The effects of technical assistance

The provision of technical assistance was one of the two core activities undertaken by the AMNEP. Technical assistance could be provided at the planning, implementation or evaluation stages of a project. Examples include the development of a partnership framework for NGOs working with DFAT in Laos and the provision of 'Ending Violence Against Women Partnership Training and Workshop' in Cambodia. The clients were typically national and international NGOs in projects funded by DFAT, either as stand-alone projects or as components within DFAT projects.

All technical assistance provided by the AMNEP was immediately evaluated by the clients of that assistance, using Adviser Assessment Forms. The usefulness of these ratings to analyse what kind of technical assistance makes a difference was very limited because, although there was an adequate rating scale, the range of ratings that are actually used was very small. This appears to have been because the use of lower ratings would affect the payment of consultants contracted by the AMNEP and this in turn would cause administrative bottlenecks in implementation of subsequent activities.

An alternative assessment method was tested by asking the AMNEP staff to rate each completed technical assistance package on multiple OECD-DAC evaluation criteria, via a card-sorting process. The results highlighted a wider range of differences between projects while still showing significant agreements between the participants in the ratings they gave.

However, as a 'supplier' perspective, it lacked sufficient independence and ideally would be complemented by a user's perspective.

A third assessment method involved an online survey of the AMNEP's technical assistance clients, which asked about immediate and longer-term (expected) effects. In both sections there were facet questions based on QAI criteria[14] plus a global satisfaction question. This instrument was used to assess client views just before the closure of the AMNEP in late 2015 (more information hereafter). This survey instrument is another example of a loose theory, in that there were no predefined views on how responses to the individual facet questions would relate to those to the global satisfaction questions; those relationships would only become known through analysis of survey responses. The ways in which this type of analysis can be done are described in the following sections.

The effects of policy dialogues

Since early 2013 there have been four AMNEP forums that have invited both NGOs and AusAID (DFAT) staff to hear and exchange views on particular development issues. Short survey instruments have been used during each of the last three forums, with two purposes in mind:

[12] These are terms used in job satisfaction literature. See Steger, M. F., Dik, B. J., and Shim, Y. (in press). Assessing Meaning and Satisfaction at Work. www.michaelfsteger.com/wp-content/uploads/2012/08/Steger-Dik-Shim-assessing-PP-chapter-in-press.pdf

[13] Four individually assessed attributes of a relationship generate a space of 24=16 possibilities, versus only four if they are arranged in a linear scale.

[14] Relevance, efficiency, effectiveness, gender equality, sustainability.

The first was to gather immediate feedback on the perceived value of the event, in order to inform the design of subsequent events. A weakness in the design of the feedback survey for this purpose was the failure to include a global satisfaction question, in addition to the specific questions about individual sessions. This meant that it was not immediately possible to identify which session(s) contributed the most to participants' overall satisfaction with the event. However, in these circumstances an aggregate score, summarising ratings given across all sessions, can be calculated in its place as the next best alternative. This is in effect another form of loose theory in that the same aggregate score can be achieved by multiple different combinations of session rating scores.

The second purpose was to collect participants' opinions about what sort of follow up activities they would do after the event. The more these were specified the more it might be possible to make follow up inquiries about the short- to medium-term effects of the events. With each session participants were asked what people and organisations they would like to make contact with thereafter. When combined with basic data on who had participated in which particular sessions, this information was used to do a simple form of social network analysis. Of particular interest was the extent to which NGO staff nominated DFAT staff or other NGO staff, and vice versa.[15]

3.6. Documenting AMEP outputs

A core feature of the AMNEP M&E framework was the use of a relational database. The database (an automated management information system, known as AMIS) was subsequently developed in 2014 using a web-accessible version of Microsoft Access and was used to store data on all AMNEP activities and outputs as they took place. In addition it was intended to include data on short- and longer-term outcomes as captured by the various instruments described above. The immediate purpose of the database was to facilitate progress reporting of the range of AMNEP activities, which it did so efficiently. The longer-term objective was to enable analysis of the relationships between AMNEPs outputs and short- and longer-term outcomes.

A database of some kind is an essential requirement where loose theories of change are being used, so as to capture and make available all data necessary for analysing all possible combinations of causes that might be at work at any given levels of investigation. For example, finding out which ratings on which forum sessions made the biggest difference to overall satisfaction levels. Or finding out what aspects of AMNEP assistance were associated with higher levels of client satisfaction.

3.7. Analysing the available data

Which relationships to examine

By 2014 a revised and much simplified theory of change was developed and summarised diagrammatically (Figure 2). The links between events indicate in broad terms the kinds of causal relationships that need to be investigated to identify what works in what circumstances. In particular:

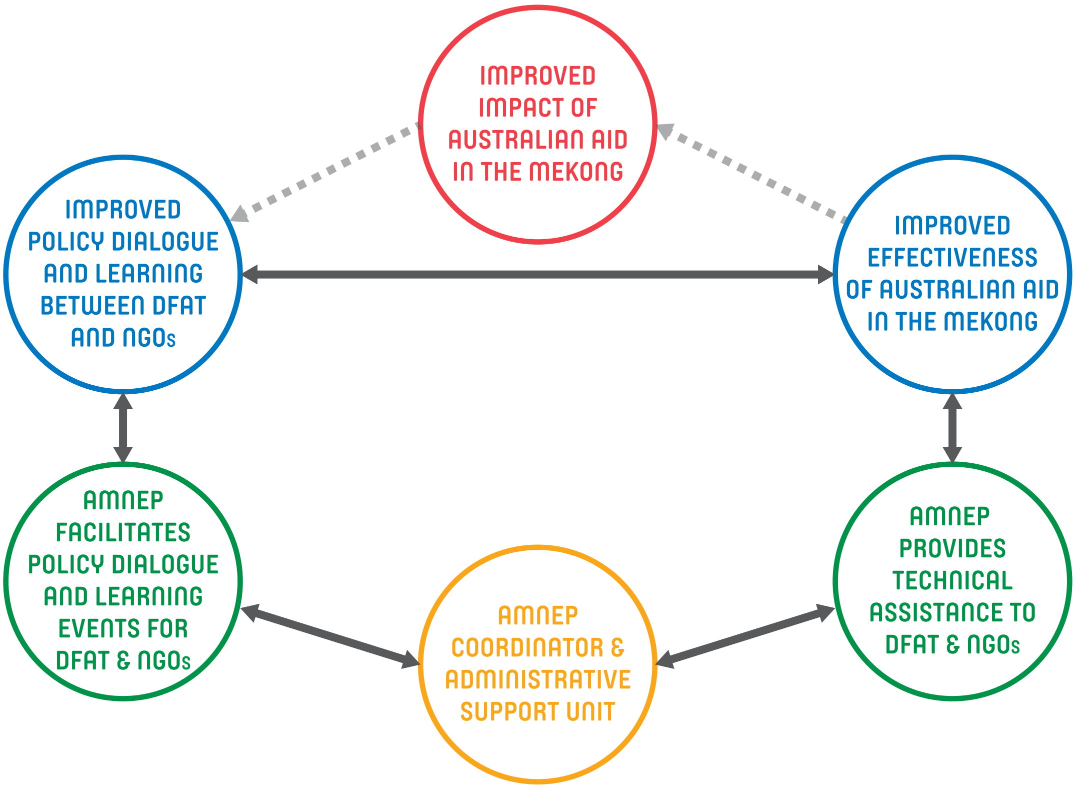
- the effects on DFAT-NGOs relationships of AMNEP facilitated policy dialogues and learning events
- the effects on aid effectiveness of AMNEP assistance to DFAT-funded NGO projects.

In addition, there were some expected interaction effects: the effects of improved aid effectiveness on DFAT-NGO working relationships and vice versa.

As already suggested, in practice the analysis that was needed was not between AMNEP outputs and longer-term outcomes per se but between participant/client responses to AMNEP outputs and longer-term outcomes (i.e. changes in the status of relationships between DFAT and NGOs, and changes in aid effectiveness).

[15] Relative to the absolute numbers of these as participants in the workshop.

Figure 2: AMNEP's 2014 theory of change


IMPROVED IMPACT OF AUSTRALIAN AID IN THE MEKONG
IMPROVED POLICY DIALOGUE AND LEARNING BETWEEN DFAT AND NGOs
IMPROVED EFFECTIVENESS OF AUSTRALIAN AID IN THE MEKONG
AMNEP FACILITATES POLICY DIALOGUE AND LEARNING EVENTS FOR DFAT & NGOs
AMNEP COORDINATOR & ADMINISTRATIVE SUPPORT UNIT
AMNEP PROVIDES TECHNICAL ASSISTANCE TO DFAT & NGOs

How to analyse causal relationships

The first AMNEP M&E report outlined an approach to the analysis of data that was informed by recent uses of Qualitative Comparative Analysis (QCA) for evaluation purposes (Befani 2013), specifically to identify the different casual configurations that may lead to an outcome of particular interest. Since then the approach that was proposed has been articulated in more detail, and now includes a wider range of options for analysis, which are discussed in the following sections.

In summary, there are three stages within the proposed overall approach, all of which are relevant to projects like the AMNEP, which aim to be flexible and responsive:

1. Search for attributes of AMNEP assistance, and the assisted projects, that *could* make a difference to outcomes of interest. This is what took place as various measurement options were considered above, especially in the design of measures that embodied loose theories.
2. Search for associations. Configurations in the collected data of specific attributes associated with the outcomes. By mid-2015 this had not yet taken place but a range of methods of doing so were available.
3. Search within cases[16] with the same configurations for any common underlying causal mechanisms. Being a more time-consuming process, this would be most feasible during mid-term and end of project evaluations. This had not yet taken place in the AMNEP but a strategy for taking it forward had been identified.

Each of these stages are described herein.

1. The search for relevant attributes

The design of survey instruments, such as those used to assess the nature of DFAT-NGO working relationships or clients' views of AMNEP assistance, involves making choices about which attributes to collect data on, given what could make a difference. For instance: the perceived competence of the other party in the DFAT-NGO relationship or the relevance of the assistance provided by the AMNEP. Those attributes that are chosen make up the loose theory that is embedded in these survey instruments. Those choices are then embedded as data fields in the AMIS database and become part of the set of attributes available for analysis.

Of equal importance are an additional set of attributes that arise from AMNEP staff conjectures about aspects of their assistance and/or the projects they are assisting (i.e. their context) that could make a difference to the two main outcomes of concern (improved aid effectiveness and improved relationships between DFAT and NGOs). Some of these attributes were identified during the initial design of the database and others during a card-sorting exercise used to identify key differences between the different forms of technical assistance that had been provided to date, in mid-2014. They included:

[16] Here cases might be different packages of AMNEP assistance provided to DFAT-funded NGO projects, as documented in separate task notes.

With regard to improved aid effectiveness:

- the stage on the project cycle when AMNEP technical assistance was provided: earlier inputs *might* be expected to have a bigger effect
- the scale of the input relative to the scale of the project: larger-scale inputs *might* be expected to give larger effects
- numbers of inputs per project: repeat requests *may* reflect the value placed on previous inputs
- the selection of the technical assistance consultant: selection by the client programme *may* improve the relevance and effectiveness of their work
- repeat use of the same consultant: this *may* lead to a larger and more sustained impact of the technical assistance
- flexible 'help desk' type terms of reference for support that is provided: this *may* improve the relevance and effectiveness of AMNEP support.

With regard to improved working relationships:

- the identities of NGOs who are repeat participants in AMNEP facilitated policy dialogues and learning events: NGO representatives who repeatedly participate in these events *might* be expected to have better relationships with DFAT
- the identities of NGO participants who identify DFAT staff as people they want to make contact with after these events: these *might* be expected to have better relationships with DFAT (and vice versa)
- the identities of those who participate as presenters: these *might* be more engaged with other participants and this may positively influence their relationships with DFAT or NGOs
- the content of forum presentations: co-participation in particular forum sessions *may* influence relationships between DFAT and NGO staff
- co-participation frequency across all events: NGO agencies and DFAT staff who have the highest levels of co-participation in AMNEP events *may* have improved working relationships

Both of these lists make up a loose theory of what might make a difference to the respective outcomes.

2. The search for associations

A search for associations is a pattern-finding activity, one that is based on an important assumption that associations between events are a *necessary but insufficient* basis on which to make plausible claims about causation. This search process is needed because of the combinatorial complexity of loose theories of change: if there are ten facets of a relationship between DFAT and an NGO that could be contributing to the overall level of satisfaction in that relationship, as expressed by the respondents, this means there are 2^{10} or 1,024 different combinations of these that could be predictors of that satisfaction level. The more attributes of a relationship that we inquire about, the more the number of possibilities to be explored expands, exponentially. In these circumstances a hypothesis-led approach that conjectures that if A+B+C facets are present then satisfaction will be high is quite a big gamble. It is one of many possibilities and to only test this one would be to take quite a gamble on their being no other alternative configurations that also work.

The alternative is to use algorithms. These are automated systemic search processes that look for and test the available combinations of attributes to determine what the best predictors of the outcome of interest are.

At time of the development of the AMNEP M&E framework there were two well-known types of algorithms that could help find potentially meaningful associations between attributes of AMNEP assistance and outcomes it was concerned with. The first is the Quine-McCluskey algorithm, built into software used for QCA.[17] QCA was first developed and used in the field of political science but is now being used in evaluations. The second type is those algorithms used to produce Decision Tree models in the field, known as predictive analytics.[18] The latter has the advantage of producing results that are much easy to communicate to and be understood by others.

More recently the author has developed a simple Microsoft Excel application (known as EvalC3) that enables the use of two other kinds of search algorithms: evolutionary search using an Excel add-in known as Solver[19] and exhaustive search, which explores every possible combination of attributes in a loose theory. The Excel application also allows very quick manual exploration, i.e. testing of specific hypotheses, which is important if participatory forms of search are a high priority.

The value of this wide range of search options is that the results generated by one method can be checked against those of another, while all use the same set of data. There is a common framework for making such an assessment, based on the use of a type of truth table, known as a 'Confusion matrix'.[20] This kind of opportunity is unusual in the evaluation of development projects.

[17] https://en.wikipedia.org/wiki/Quine%E2%80%93McCluskey_algorithm

[18] https://en.wikipedia.org/wiki/Predictive_analytics

[19] www.excel-easy.com/data-analysis/solver.html

[20] https://en.wikipedia.org/wiki/Confusion_matrix

3. The search for causal mechanisms

Such an inquiry is most appropriate during mid-term or end-of-project evaluations, when there is dedicated time available. Systematic searches of the kind described herein tell us, in effect, where to focus our attention in a large haystack of possibilities. However, while systematic search processes can provide us with associations that are good predictors of outcomes, we don't yet know if there is any causal mechanism at work underlying these associations. To answer this question we need to move from cross-case comparisons to within-case investigations.

There are two steps involved here, neither of which had yet been implemented by the time the AMNEP was prematurely closed down. The first step is the selection of appropriate exemplar cases, and the second, in-depth inquiry within these. Three types of exemplar cases[21] are likely to be of value:

- NGO projects where a configuration of attributes of AMNEP assistance is associated with improved aid effectiveness (known as 'true positives'). Inquiries here will look here for a plausible causal mechanism at work
- NGO projects where the same configuration of attributes is associated with no improved aid effectiveness (known as 'false positives'). Inquiries here will look for the presence of the same causal mechanism but some other factors that prevent it from working
- NGO projects where the absence of the same configuration of attributes is associated with improved aid effectiveness (known as 'false negatives'). Here it is expected that the casual mechanisms found will not be present but one or more other mechanism, associated with other configurations of AMNEP assistance, will be.

[21] A more detailed explanation of complimentary within-case inquiries can be found here: http://evalc3.net/how-it-works/within-case-analysis.

4. Reality check: what was actually possible?

4.1. AMNEP's contribution to improved aid effectiveness

By late 2015 QAI data was available for 2 or more consecutive years on 7 of the 13 AMNEP assisted projects documented in the AMIS database. The same kind of data was also available on 69 other DFAT projects in the same countries. As might have been expected, AMNEP-assisted projects varied in the extent to which they improved over time or not: four improved, two declined and one remained the same.[22] Relative to the wider set of DFAT-funded projects their performance was similar, with three doing better than average, three doing worse and one the same. Had AMNEP continued until late-2018 or beyond, the number of projects with this kind of data would have been larger.[23]

In order to assess AMNEP's contribution to changes in aid effectiveness at least one additional set of data would also be needed, i.e. data about client views of the value of assistance provided by AMNEP. In late 2015, 18 clients of AMNEP assistance were asked, via an online survey, for their opinion on the immediate and longer-term value of the assistance that had been provided by the AMNEP to specific projects they had been engaged with.[24] At the time of writing only a small proportion (4 of the 18) had responded and of these only two were about projects for which there was also data on changes in QAI ratings over time. It is not possible to do a cross-case analysis with this number of cases, using the methods described in section 3.6.2 above. Had AMNEP continued as planned it is likely the number of cases would have grown and the proposed analysis would have been possible.

Assuming that AMNEP support was causally linked to changes in aid effectiveness it is clear that a statement about the average net effect of AMNEP support would not be very informative. On average it seemed to have made no difference but beneath this average there was substantial variation. The alternative approach, as proposed in the first M&E framework report on the AMNEP, was to think in terms of multiple, conjunctural causation. In other words, there may be multiple casual pathways involved, some leading to improvement in aid effectiveness and some to a reduction. Each of these may involve combinations of different attributes, both of the assistance provide and the kind of project assisted. If the AMNEPs evaluation objective is to find out what works in what circumstances, then these different pathways need to be identified. That would be possible using the kinds of algorithms described in section 3.6.2, if sufficient data were available.

4.2. AMNEP's contribution to improved relationships between DFAT and NGOs

As of late 2015 no data had yet been collected on DFAT and NGO staff's view of their working relationships with each other, although a survey instrument had been developed.

Lack of progress with measurement of the second objective suggests there are limits as to how 'loose' a useful loose theory of change can be. In the analysis of improved aid effectiveness there was always a clearly indefinable list of projects, assisted by the AMNEP, whose aid effectiveness was under scrutiny. But with the second objective of improved working relationships between DFAT and NGOs, while there was a usable measurement instrument, it had not been clearly defined whose relationships would be evaluated and whose would not. The theory of change suggested these would be those that attended AMNEP-facilitated forums and learning events. Data was collected on these participants but their membership varied substantially from event to event, with relatively little overlap, which raised questions as to how widely the second objective applied to.

The theory of change shown in Figure 2 also suggested that participants in projects assisted by the AMNEP would have their relationships affected by that experience. But, by mid-2015, the AMIS database was not sufficiently developed to document names of people engaged via this channel.

The reasons behind the lack of clarification of the expected target group are probably not just technical: canvassing NGOs opinions of DFAT may have been seen as problematic during a period when DFAT was both absorbing AusAID and cutting back the size of the aid programme. In a period of uncertainty, simply being able to deliver some useful meeting opportunities may have been seen as an achievement in its own right.

[22] QAI/AQC scores for each project in each year were aggregated, then changes calculated between years and then reduced to a common per annum rate

[23] And the duration of the time series would have been longer.

[24] The scale of this survey was limited to the requestors of AMNEP assistance, who were all AusAID/DFAT staff. The survey did not include staff of NGOs whose organisations involved with these projects.

5. Lessons to be learned

5.1. The search for relevant attributes

On reflection, the process of identifying conjectures about relevant attributes of project (and AMNEP assistance to these) and converting them into data fields and data in AMIS was not sufficiently systematic. In practice, it took place at two points in time: during an initial ethnographic exploration of types of technical assistance; and, later, at the time of database design. Ideally this should have been done more periodically to capture within AMIS updated understandings of how AMNEP's activities were having an impact. Additionally, these conjectured fields should be visibly tagged as such and distinguished from other data fields in AMIS used for periodic reporting purposes, both to facilitate analysis and ensure they are receiving separate, ongoing attention.

5.2. The search for associations

The ability to analyse different causes of impact (or lack thereof) is dependent on the number of cases available for analysis, e.g., in this case, AMNEP-assisted projects. For example, if there are 8 cases, then these could illustrate, at best, all the possible combinations of three attributes of those projects and any assisted provided (i.e. 2^3). For 4 attributes, 16 cases would be needed, if all combinations are realistic possibilities (i.e. 2^4). If there are many attributes and few cases there are two risks. One is that more than one association could be found that fits the data. The other is that any association found within those cases may be disproved by any new cases with previously unseen configurations of attributes. These problems are not new to users of methods like QCA or decision tree algorithms, and there are both data- and theory-oriented approaches to dealing with them.[25] However, in a context such as that of the AMNEP, the appropriate response is to do some more careful thinking at the stage of identifying and selecting attributes to collect data on; having a 'loose theory' does not mean having no theory.

5.3. Short cycle learning

Impact evaluation is typically seen as a process that takes place over the longer term, once sufficient data has been collected and the effects of project interventions have had their opportunity to play out. But impact evaluation in the sense of uncovering causal configurations at work does not necessarily have to be this limited. Causal processes can work at many different temporal scales. This view was the basis for the proposal to look for changes in project QAI ratings, rather than changes in the lives of people targeted by those projects. There was also a possibility of analysing much shorter-term changes. Two of the survey instruments developed for the AMNEP were structured in such a way that they contained loose theories about possible determinants of client satisfaction – either with AMNEP assistance or with DFAT-NGO relationships. Both the outcomes and determinants may well be subject to change in the short term, and be worth tracking and analysing. If so, then perhaps we should also be discussing impact monitoring, not just impact evaluation.

5.4. A wider view of evaluability

An initial assessment of the AMNEP made in 2013 was that it was unevaluable. However, with hindsight one of the initial grounds for the M&E consultant's judgment of unevaluability was misconceived, namely the 'lack of identifiable (and thus testable) causal linkages between the 50 or more events in the much more detailed diagrammatic theory of change that was developed shortly after project approval'. First, this is not always a realistic expectation: there are settings where developing such a clearly articulated theory of change is neither possible nor desirable. Humanitarian emergencies and participatory development projects are two archetypal situations.

Second, and equally important, as introduced in this paper there are now means of systematically and transparently identifying linkages between events and outcomes after the fact. Loose theories of change are needed and can be worked with. But this added flexibility comes with some conditions. Data needs to be collected, as is already the case with a tight theory of change, and probably more data than before, because in a loose theory of change it is likely that there will be a wider range of interventions and outcomes where relationships are possible and therefore need analysis. And a theory is still needed as a basis for screening out what might be possibly relevant types of data, or not. This kind of theorising will need to be more divergent than convergent, open to the possibility of a diversity of casual pathways – and might help bring us closer to the real world.

[25] In both areas the problem has a specific name: the problem of limited diversity and the curse of dimensionality, respectively.

References

AusAID (2012) Australia Mekong – Non-Government Organisation Engagement Platform. Final design document.) https://dfat.gov.au/about-us/publications/Documents/mekong-ngo-engagement-platform-design-doc.pdf

Befani, B. (2013) Between complexity and generalization: Addressing evaluation challenges with QCA. Evaluation July 2013 Vol. 19 No. 3 269-283.

Davies, R. (2013) Planning Evaluability Assessments: A Synthesis of the Literature with Recommendations. Report of a Study Commissioned by the Department for International Development. DFID.

Eyben, R., Guijt, I., Roche, C. and Shutt, C. (2015) The Politics of Evidence and Results in International Development: Playing the Game to Change the Rules? Practical Action Publishing.

Peersman, G., Guijt, I., Pasanen, T. (2015) Evaluability Assessment for Impact Evaluation. Guidance, checklists, and Decision Support. Overseas Development Institute. London.

Ragin, C.C. (1989) The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies. University of California Press.

Ramalingam, B. (2013) Aid on the Edge of Chaos. Rethinking International Cooperation in a Complex World. Oxford University Press.

Siegel, E. (2013) Predictive Analytics. John Wiley & Sons.

Snowden, D. (2015) Cognitive Edge blog. http://cognitive-edge.com/blog/author/dave-snowden/

Vowles, O. (2015) Adaptive Programming. Posting on DFID Bloggers - https://dfid.blog.gov.uk/2013/10/21/adaptive-programming.

ODI is the UK's leading independent think tank on international development and humanitarian issues.

Readers are encouraged to reproduce Methods Lab material for their own publications, as long as they are not being sold commercially. As copyright holder, ODI requests due acknowledgement and a copy of the publication. For online use, we ask readers to link to the original resource on the ODI website. The views presented in this paper are those of the author(s) and do not necessarily represent the views of ODI, the Australian Department of Foreign Affairs and Trade (DFAT) and BetterEvaluation.

ISSN: 2052-7209
ISBN: 978-0-9941522-1-3

All ODI Reports are available from www.odi.org

Overseas Development Institute
203 Blackfriars Road
London SE1 8NJ
Tel +44 (0)20 7922 0300
Fax +44 (0)20 7922 0399

www.odi.org