ALNAP

ALNAP Discussion Series

Improving the quality of EHA evidence

Method Note 4

January 2015

Addressing causation in humanitarian evaluation: A discussion on designs, approaches and examples

by Jessica Alexander and Francesca Bonino

Introduction

Programme managers, donors, decision makers at all levels want to answer the question – *did our intervention lead to the desired impact? What difference did our intervention make?* [1]

Understanding cause-and-effect, or causation, is what distinguishes evaluations from other processes such as assessment and monitoring. However not all evaluations address, or are expected to answer questions about cause-and-effect (see **box 1**). Many observers have noted that most evaluative work focuses on process issues, describing what happened during an intervention and how, rather than answering the question of whether the intervention brought about the intended results (outcomes and impacts)(Morra Imas and Rist, 2009:228).

Hughes and Hutchingson from Oxfam GB summed it up well:

> 'At the moment, much of our evaluative efforts are spent on evaluation designs that focus on establishing whether intended outcomes have materialised rather than on assessing the contribution of our interventions to such changes.' (2012: 10)

Yet over the last decade, humanitarians have increasingly tried to tackle the harder and more complex question of causality. As Knox-Clarke and Darcy (2014) noted in a study on quality and use of evidence in the humanitarian sector, 'It is not enough for an evaluation to depict a situation accurately; it also needs to show the relationship between a specific intervention, or series of interventions, and the situation described.' (p.40)

This note explores this issue, offers methods used and examples of establishing causality in evaluation of humanitarian action (EHA) work.

1 The authors would like to thank John Cosgrave for his helpful comments to an earlier draft of this note.

Box 1: Causation: One of four main types of questions addressed in an evaluation

There are four main types of questions that can be addressed in an evaluation.

1. **Descriptive questions** – may describe aspect of a process, a situation, a set of organisational relationships or networks; can be used to describe inputs, activities and outputs.
2. **Relational questions** – they are very common in evaluation and involve establishing whether a relationship between two or more phenomena exists at all, and if yes, its direction and magnitude.
3. **Normative questions** – they compare 'what is' with 'what should be' and the current situation with a specific target, goal, or benchmark.
4. **Causation questions** – determine what difference the intervention makes. They are questions about establishing cause-and-effect on an intervention and ask whether the desired results have been achieved as a result of a programme/intervention or by other factors.

Many evaluations ask only descriptive and normative questions, particularly if they are formative evaluations that focus on implementation of an intervention. Evaluations focusing on outcomes and impacts ask cause-and-effect questions, but they typically also include some descriptive and normative questions (Morra Imas and Rist, 2009: 228).

Source: Morra Imas and Rist, 2009:223-229; and Coryn, 2013

Framing the issue

Causation in evaluation refers to the process by which evaluators identify and document a causal relationship between a particular intervention and change in the conditions of people's lives and context. This means that evaluators need to be able to document that a given result, change, or effect has been *caused* by the intervention and not by coincidence. Morra Imas and Rist (2009: 227) offer this example:

> Take an intervention that introduced farmers to a new and improved seed. A descriptive evaluation question could ask whether the intended outcome has been achieved. The evaluation could ask whether the grain yield increased and if yes, by how much. However, decision-makers and the evaluation commissioning agency may also be interested to answer a different evaluation question that asks whether the crop increased as a result of the intervention – and not, for instance as a result of unusually ideal weather for the grain crop. This is a clear example of a cause-and-effect question.

While the above example comes from a development context, a humanitarian programme may for example want to know to what extent was the observed reduction in malnutrition due to the supplementary feeding programme?

Establishing this causal relationship (also called **causal inference**) is not straightforward. It may not be possible to isolate the results brought about by a given intervention amongst a host of other factors at play in a given context. In the example above looking at reduced malnutrition, there could be a range of factors at play contributing to the observed results such as: other concurrent health-related interventions (such as deworming); improved hygiene conditions; seasonal factors; and higher household income.

This points to what is commonly referred to in evaluation as the ***attribution problem***, and is the focus of an extensive body of literature in the evaluation field. Attribution requires establishing the causal implications of an intervention and/or the causation of a phenomenon that can be observed. (Scriven, 2010: i). (For an introduction on this specific issues see for instance Gerring, 2012). However, establishing attribution does not equate with establishing 'sole attribution' and some have addressed this by exploring ***contributory causation*** as discussed below.

Establishing causation is a challenge for evaluators across the many branches of evaluation practice – but within the humanitarian sector, there are even greater complications when addressing cause-and-effect questions. Given the rapidly-changing and often unstable environments where humanitarian actors operate, designing and carrying out such causal analysis can be daunting as there will inevitably be different types of biases[2] which impact data collection and analysis. Other realities which may limit causal analysis stem from:

- limited high quality performance data (including monitoring data at different levels of results including at outcome level)
- scarce availability of skills and expertise to design and carry out causal analysis in small n situations – (situations when there are too few units of assignment to permit tests of statistical difference in outcomes between the treatment group and a properly constructed comparison group)(see Gerring 2012; and White and Phillips, 2012)
- tendency to not plan for evaluations early enough in the programme cycle, minimising chances that periodic assessments and monitoring data will support a causal analysis at a later evaluation stage.

"Theory-based evaluation should not be seen simply as a replacement for experimental and quasi-experimental designs."

Nevertheless, demand is growing to find ways to overcome these constraints which are discussed in a recent paper by 3ie – the International Initiative for Impact Evaluation (Puri et al., 2014) and which this post touches on below.

Before delving in the specifics of different approaches and tactics to address causation in evaluation (later in section II) it is useful to provide the broad backdrop against which those approaches situate. Table 1 below gives an overview on the **three main families of evaluation designs**[3] and displays what makes them different when we opt to use them as basis to infer causation.

2 See for instance a previous EHA method notes that discussed representativeness and accuracy of evaluative evidence in humanitarian settings. Alexander and Cosgrave (2012); Alexander (2012).
3 An evaluation design is the overarching logic of how we conduct an evaluation. Different designs can be used to answer different types of evaluation questions.

You may use different designs to answer different evaluation questions, but typically one type of design will predominate.

Table 1: 'Families' of evaluation design and requirements for dealing with causation

Broad categories of design (or families of designs)	Possible options	Some characteristics	Best used for which type of intervention (Stern, 2008)	On which basis do they infer causation?
Experimental	RCTs Natural experiements (*)	Strongest design option to control for selection bias in evaluation because the subjects are randomly assigned to the intervention.	Standardised interventions in identical settings with common beneficiaries.	Random assignment of the intervention to groups.
Quasi-experimental	Before-and-after designs without comparison group; Interrupted time-series; Longitudinal designs; Panel design; Correlational design using statistical control; Regression discontinuity designs	They make no use of randomisation. There are broad sub-groups of quasi-experimental designs. Those using **matching designs** (with various constructed groups) and those using statistical designs. While some comparison designs are statistically based (for matching) not all **statistical designs** involved constructed groups. For example in Interrupted Time Series, the whole population is sample, before and after the key event. The evaluators 'construct' groups that are as equivalent on important characteristics (gender, income, socio-economic background) as possible. Sometimes the evaluator can create a comparison group by matching key characteristics. For example when large samples are used, or good secondary data is available, it is possible to use statistical matching techniques such as Propensity Score Matching. Careful matching of treatment and comparison group can eliminate or greatly reduce the likelihood that rival explanations exists for a given result. Such rival explanation could also be that the two group were different from the start – and those different features are those that explain what brought about a result. (Davidson, 2005: 246)	Standardised interventions in diverse settings, possibly with diverse beneficiaries	Establishing comparison groups and / or Repeated measurement over time and /or Before / after comparisons

Broad categories of design (or famlies of designs)	Possible options	Some characteristics	Best use for which type of intervention (Stern, 2008)	On which basis do they infer causation?
Non-experimental (or descriptive, relational)	Theory-based designs for example Process Tracing; Theory of Change and impact pathways; Contribution Analysis; Realist Evaluation analysis	No attempt is made to establish intervention and non-intervention groups.	Customised interventions in diverse settings with diverse beneficiaries that use narrative/qualitative approaches to build plausible explanation of results	Identification / confirmation of causal processes or 'chains'. Identification and confirmation of supporting factors and causal mechanisms at work in a given context.
	Case-based designs for example Grounded theory approaches; Ethnography; Qualitative Comparative Analysis (QCA); Within-case analysis; Network analysis	**Non-experimental designs provide an in-depth description of a phenomenon or the relationships between two or more phenomena (OIOS-IED, 2014: 49).**		**Comparisions across and within cases. Analytic generalisation based on theory.**
	Participatory designs can use for example Empowerment approaches; and Collaborative Action Research approaches	The emphasis is on describing the size and direction of a relationship between different observed outcomes and explaining the causal mechanisms at work in a given context.		Validation by participants that their actions and what they experience have been brought about by a given programme/ intervention.
	Synthesis designs for example Narrative Synthesis and Realist based synthesis			**Accumulation and aggregation within a number of different perspectives (statistical, theory-based, ethnographic etc).**

(*) It is debatable whether natural experiments should be grouped within the 'family' of experimental designs, and hence considered as 'true experiments' because the comparison group is created by chance rather than by random allocation.

Adapted from: Morra Imas and Rist, 2009:249-280; Stern, 2008, Stern et al., 2012:15-35; Bamberger, Rugh, Mabry, 2012:213-244.

II. Approaches to addressing causation in evaluation

This section gives an overview on different approaches that can be used to establish causation and while not meant to be exhaustive, offers some of the most frequently used in the broad field of programme evaluation. [4]

Randomisation

In a randomised control trial (RCT) participants are **randomly assigned to a treatment (the group that received the intervention) or control group** (the group that did not receive the intervention or received an alternative intervention).

"The depth and breadth of the required evidence base is a key consideration in evaluation planning"

Provided that sampling is done carefully and that sample sizes are large enough, randomisation helps ensure that there are no systematic differences between the group that received the intervention and those who did not. This means that any observed differences can be casually attributed to the intervention (Davidson, 2005: 240).

While considered by some a so-called 'gold-standard' in establishing causality, these approaches can be time consuming and resource-intense because of the large sample size required to ensure that confounding factors are evenly distributed between 'treatment' and 'control' groups. (For more on these points, see for instance, Jones, 2009).

Evaluator insight 1: Dean Karlan warns about the misperception of RCTs

> A common misperception is that one must choose either to do a qualitative evaluation or an RCT. Underlying this is an erroneous spectrum of 'attribution' rigor, with RCTs on one end and qualitative methods on the other. In reality, qualitative methodologies are not the opposite of RCTs. For one, a good RCT evaluation often involves a thorough assessment of how the program functions, its initial design, theory of change, beneficiary participation, etc. (Karlan, 2009 pp.8-9)

The evaluator insight highlights that the choice of evaluation design does not determine the choice of methods (predominantly qualitative, quantitative or mixed-method) used to gather and analyse the data.[5] It is an important point, especially in light of the often polarised commentaries proposed by evaluators and researchers following different evaluation traditions and method orientations. (For a comprehensive discussion on this specific point see: Goertz and Mahoney, 2012)

In their study on *Improving Humanitarian Impact Assessment: Bridging theory and Practice* Proudlock, Ramalingam and Sandison (2009) point out despite the ethical concerns of RCTs, random designs are best at eliminating selection bias and at minimising confounding factors.
Randomised designs can also be particularly useful in providing insights into the benefits of certain interventions over others – such as between cash or food aid.

4 See BetterEvaluation site for a longer menu of options: http://betterevaluation.org/plan/understandcauses and also http://betterevaluation.org/themes/impact_evaluation
5 John Cosgrave - Personal communication with the Authors – 12 December, 2014.

However, findings from randomised designs need to be backed up by significant contextual analysis because it is impossible to tell how much the specific context of an RCT shapes the outcomes and impacts observed. In their study Proudlock, Ramalingam and Sandison (2009) cite a statement by the European Evaluation Society (2008) which proposed that RCTs should only be considered in cases where:

- a linear causal relationship can be established between the intervention and desired outcome
- it is possible to 'control' for context and other factors
- it can be anticipated that programmes under both experimental and control conditions will remain static for a considerable time
- randomness and chance has no influence
- it is ethically appropriate to engage in randomisation.

The recently-published 3ie working paper (Puri et al., 2014) on impact evaluation in the humanitarian sector touches on issues of ethical viability of impact evaluation designs that use randomisation or a constructed control group. The study also features several case studies of how different humanitarian agencies have addressed ethical, design and data challenges to arrive at robust statements of causal linkages.

To conclude, Tom Cook reminds us that random assignment cannot by itself guarantee a secure causal inference nor that results are not skewed if 'blind' and 'double blind' designs where not used. Cook advises researchers and evaluators opting for randomised designs to:

- justify that a correct random assignment procedure was chosen
- document that the random assignment procedure was implemented well
- show that attrition was not differential by treatment
- show that treatment contamination was minimal or otherwise adequately dealt with. (Cook, 2010: 111)

Establishing counter-factuals

Counterfactuals are often used when establishing causation. This approach seeks to answer the question, 'What would have happened without the intervention?' by comparing an observable world with a theoretical one, where the latter is intended to be identical to the former except for the presence of the cause and effect (DFID, 2011).

This is not an experimental method, merely a comparison of two similar situations, one where an effect and an event supposedly representing its cause has taken place, and the same situation where the effect and cause have not taken place. If two such events can be shown to exist, the effect is attributed to the cause. But the assumptions of counterfactual thinking do not always hold and it may be impossible to find an identical match for the factual world. Even when this is possible, counterfactuals associate a single cause with a given effect without providing information on how the effect is produced. Without knowing the how, evaluators may miss out on critical information (DFID, 2011).

Establishing the causal package

"The evaluation questions should drive method and design choices. Not the other way around."

Causality, which involves relationships between events or conditions, is often discussed in terms of necessary and sufficient conditions (Mayne, 2012). In most cases we assume that there are numerous factors at play aside from the intervention which make it impossible to conclude that it was sufficient to 'cause' a result. Yet, we can expect that our intervention, along with other influencing factors (what Mayne refers to as a 'causal package'), is indeed sufficient to assume a causal outcome. (Mayne, 2012: 1). Mayne thus suggested that these causes together, which are neither necessary nor sufficient, can be called **contributory causes** (Mayne, 2012: 2). From this 'contributory' perspective, appropriate evaluation questions would be:

- Was the causal package of the intervention plus its supporting factors sufficient to produce the intended result?
- Is it likely that the intervention has made a difference?
  - Is it likely that the intervention was a contributory cause of the result?
  - What role did the intervention play in the causal package?
- How and why has the intervention made a difference?
  - How did the causal factors combine to bring about the result?
  - What context was relevant and which mechanisms were at work?
- Has the intervention resulted in any unintended effects? (Mayne, 2012: 1-2)

One recent example of evaluation methodology that focuses entirely on understanding contribution has been developed by the University of East Anglia and Oxfam GB (Few et al., 2014). This approach – called Contribution to Change – has been presented by its proponents as best suited to capture and describe the relative importance of post-disaster interventions in aiding people's recovery as opposed to focusing on establishing specific changes due to single agency interventions (p.8). Furthermore, the proponents of this methodology also suggest that the focus on contribution makes this methodology 'particularly appropriate for joint evaluation exercises between agencies' (p.5). One example where this methodology is currently being applied is in the context of a UK Disaster Emergency Committee (DEC) commissioned evaluation of the response of DEC's members to the Typhoon Hayan in the Philippines.[6]

Using programme theory to infer causation

Theory-based approaches (which include the use of theories of change, case studies, causal modelling, outcome mapping, most significant change and 'realist' evaluations) involve examining a particular case in depth and theorising about the underlying causal links between actions, outcomes and impacts, thus building a programme theory of change. (For an overview see for instance Rogers, 2000).

While the literature acknowledges that the findings can't be proven statistically, the approach can provide a logical argument that certain inputs will lead to a given change (Proudlock and Ramalingam, 2009) and should not necessarily be seen as a 'second best' option. Davidson (2000) offered an insightful reflection on this point – building on some work by Cook (2000):

Theory-based evaluation should not be seen simply as a replacement for experimental

6 At the time of writing the evaluation fieldwork has been completed. The evaluation report is forthcoming

and quasi-experimental designs. For high-stakes evaluations with large budgets and extended time lines, the two may be used in conjunction to allow virtually bulletproof causal attributions, provided they are used skilfully For the everyday evaluator under more serious time and budgetary constraints, ideas from both methodologies should be considered in order to build evidence for inferring causality (p.25).

Using temporal comparisons

These are comparative approaches that seek to determine differences before and after an intervention using recall methods. Before and after comparisons are also appropriate where there are no other plausible explanations for change, otherwise it may not be a sufficient measure for how the situation would have evolved in the absence of an intervention.

Interrupted Time Series (ITS) is an example of a comparative approach that makes use of temporal comparisons by providing a powerful indication of changes in the outcomes that are placed in a timeframe in line with the treatment/interventions. It is important to note that as with other temporal designs, evaluations using ITS will not establish causality but show that the variable of interest changed at the same time as the treatment.

Using regression discontinuity (RD) designs

Regression discontinuity designs can be used to demonstrate differences between a treatment and non-treatment group. They Regression discontinuity designs can be used to demonstrate differences between treatment and non-treatment group. They are an attractive option for humanitarian contexts because the treatment and control group are not/do not need to be randomly selected but are selected on the basis of some score or other metric. In practice, this means that assistance can be allocated on the basis of criteria such as vulnerability rather than on a random basis. An example of this method is highlighted in the following section.

Analysing contribution within a 'causal package' (contribution analysis)

This kind of analysis recognises that several causes can contribute to a result. Each of them – on their own – might not be necessary nor sufficient for lead to desired impact.

As John Mayne argued: an intervention on its own may not be neither necessary nor sufficient to bring about the result. We can take a different perspective on causality and still speak of the intervention making a difference in the sense that the intervention was a necessary element of a package of causal factors that together were sufficient to bring about the results. It is in this sense that we can speak of **contributory causes**. (Mayne, 2012:1)

In complex settings, causes interact unpredictably making it difficult to develop models which capture the combination of factors. Contribution Analysis emerged during epidemiological studies on the role of tobacco as a cause of lung cancer. Cancer can be caused by quite a varying mix of factors in which tobacco plays no part, such as lifestyle, environmental and genetic factors. Thus, there are other ways to get to the impact which may not include the intervention. The intervention is considered a contributory cause of the impact if:

- The causal package with the intervention was sufficient to bring about the impact and;
- the intervention was a necessary part of that causal package.

BetterEvaluation notes that the essential value of contribution analysis is that it offers an approach designed to reduce uncertainty about the contribution the intervention is making to the observed results through an increased understanding of why the observed results have occurred (or not!) and the roles played by the intervention and other internal and external factors. The report from a contribution analysis is not definitive proof, but rather provides evidence and a line of reasoning from which we can draw a plausible conclusion that, within some level of confidence, the programme has made an important contribution to the documented results.

Collaborative outcomes reporting (COR)

Collaborative outcomes reporting is a participatory, collaborative approach to understanding impact which uses story telling about how a programme contributed to results (outcomes and impacts) using multiple lines of evidence. Most are short, mention programme and context, relate to a plausible results chain, and are backed by empirical evidence (Dart and Mayne, 2005). Critical to COR is the review process by expert panels and stakeholders to check for credibility.

General elimination methodology (GEM)

General elimination methodology, involves three steps (Scriven, 2008):

1. Identifying the list of possible causes for the outcomes and impacts of interest.
2. Identifying the conditions necessary for each possible cause in the list to have an effect on outcomes or impacts.
3. Working out whether the conditions for each possible cause are present or not.

Working through these elimination steps means that the final set of causes include only those whose necessary conditions are completely present (Scriven, 2008). One of the key conditions for GEM to work is that the list of possible causes has to be exhaustive, which is unlikely in humanitarian settings or in other complex programming and implementation scenarios.

Other approaches

BetterEvaluation outlines other methods to map, explain and investigate factors which may influence programme impacts. These are:

- Force Field Analysis: providing a detailed overview of the variety of forces that may be acting on an organisational change issue.
- Process tracing: ruling out alternative explanatory variables at each step of the theory of change.
- RAPID outcomes assessment: a methodology to assess and map the contribution of a project's actions on a particular change in policy or the policy environment.
- Ruling out technical explanations: identifying and investigating possible ways that the results might reflect technical limitations rather than actual causal relationships.
- Searching for disconfirming evidence/Following up exceptions: treating data that does not fit the expected pattern not as outliers but as potential clues to other causal factors and then seeking to explain them.
- Statistically controlling for extraneous variables: collecting data on the extraneous variables, as well as the independent and dependent variables is an option for removing the influence of the variable on the study of programme results.

III. What has been tried when evaluating humanitarian action?

Not all the approaches and evaluative analytical tools mentioned in the previous section have been extensively explored (yet) in EHA practice. Nevertheless, below is a list of examples of different evaluation designs and analytical approaches used to approach the causation question in evaluation in humanitarian settings.

Example of mixed-method theory-based impact evaluation from WFP and UNHCR (2012)

WFP and UNHCR jointly commissioned and conducted a series of mixed-method impact evaluations to assess the contribution of food assistance to durable solutions in protracted refugee situations (WFP and UNHCR, 2012).

Overall, the impact evaluation series used a theory-based approach as the basis to infer causation. Moreover, a synthesis evaluation was produced using a series of four stand-alone evaluations carried out in Bangladesh, Chad, Ethiopia and Rwanda, which used the same theoretical framework and approach, but with details adapted to the context in each case.[7] Some of the key features of this impact evaluation series are its use of:

- A logic model (Theory of Change, ToC) that was first developed by WFP Evaluation Office, to be then discussed and validated by the evaluation teams in the four countries visited; the teams also assessed the match of the ToC with country-level logical frameworks at different stages of the evaluation.
- A mixed-method approach that was chosen for its best fit in a situation where it was not possible to use a design that used a conventional counterfactual as a basis to infer causation. The use of mixed-methods to gather and analyse data in the four country cases included triangulation of data generated by desk reviews; interviews with WFP and UNHCR stakeholders; reviews of secondary data; quantitative surveys; transect walks; and qualitative interviews, including with focus groups of beneficiaries and members of local refugee-hosting communities (WFP and UNHCR, 2012:2).
- An evaluative analysis drawing from the results that emerged from the country case studies to establish: a) which internal and external factors can causally explain the results; and b) which factors influenced the results and changes (intended and unintended) observed in the different countries and why (see WFP and UNHCR, 2012: 10-13).

The evaluation team explained that the series of country-based impact evaluations were used as a way of:

> 'test[ing] the validity of an intervention logic derived from the MOU [Memorandum of Understanding] between UNHCR and WFP and the two agencies' respective policies and programme guidance. This logic posited that the agencies' combined activities and inputs contributed to increased refugee self-reliance over three stages of evolution, starting from the refugees' situation on arrival. (…) All four [country-level impact] evaluations tested its assumptions and the extent to which food assistance contributed to outcome levels over time..' (WFP and UNHCR, 2012: 2)

Example of quantifying impact of cash-based assistance: IRC in Lebanon (2014)

IRC completed an evaluation of the impacts of the winter cash transfer programme run by UNHCR and partners from November 2013 to April 2014 (Lehmann and Masterson, 2014). The programme gave $575 USD via ATM cards to 87,700 registered Syrian refugees in Lebanon with the objective of keeping people warm and dry during cold winter months. This was the first research on causal-inference to understand the impact of Conditional Cash Transfers (CCTs) and of unconditional cash transfer programmes (UCTs) that focused more on long-term poverty relief and development projects, rather than solely on more

7 The link to the series of reports is available from the WFP Evaluation Office site here: http://www.wfp.org/evaluation/list?type=2711&tid_1=All&tid_2=All&tid_3=1959

immediate humanitarian assistance issues. It was also the first study to use an RCT methodology (comparing refugees receiving cash to those who did not) in order to quantify the impact of cash assistance.
This study used a regression discontinuity (RD) design that exploited the targeting strategy used as part of the cash assistance programme itself: cash transfers were made to benefit those living at high altitudes to target assistance for those living in the coldest areas during the winter months. In practice, those who met the threshold of living above 500 meters of altitude received cash and those beneath 500 meters did not. The study thus compared outcomes of beneficiaries residing *slightly above* 500 meters (group who received cash) to non-beneficiaries residing *slightly below* (group who did not receive cash), but applied the same demographic criteria to calculate vulnerability, thus limiting (as much as possible) the variation between treatment group and control group only to whether they received cash or not.

The research team went even further to check this by using UNHCR's demographic data to compare pre-treatment characteristics between the treatment and control groups. Among the demographic variables that were available, 21 of 24 variables are balanced.

Therefore, prior to the start of the programme, households in treatment and control group were very similar and had similar characteristics such as same size households prior to the start of the programme. They then assumed that 'if treatment and control group have similar characteristics prior to the start of the programme, then any difference after the start (e.g., in the April/May household survey) measures the programme's causal impact.' (Lehmann and Masterson, (2014: 6).

"Which level of certainty the evaluation commissioning agency and the intended users of the evaluation need to have in order to be confident that a given intervention led to a certain change or result?"

Example of analysing contribution: NRC evaluation of advocacy and protection in the DRC

An evaluation commissioned by Norwegian Refugee Council's (NRC) to assess results at outcome level of its 2012-13 advocacy and protection initiative in the Democratic Republic of the Congo (DRC) (O'Neil, and Goldschmid,2014) is worth noting for the following features:

- The evaluation made an attempt to **estimate the level of contribution** of the initiative to any changes seen at the outcome level (p.10) and capture unanticipated results (p.16).
- The evaluation looked at different changes: at policy level; at field level; in the area of conflict resolution; at the level of the international community; at the level of NRC specific field presence in DRC; and finally at the level of NRC programmes in general and of the NRC advocacy programme in particular.
- The analysis of NRC's contribution to change was carried using a type of rubric (presented below) to capture and map the different levels of contribution. The evaluation team focused on providing a detailed nuance of the extent to which NRC's role and activities were visible (or not) in contributing to different changes observed.

Rubric used to map the contribution to change

Change seen	Role of NRC advocacy
Unknown Evaluation was unable to assess if change had occurred (change may have occurred but we were not aware of it)	**Unknown** Evaluation was unable to assess if NRC had an influence on change (influence may have occurred but we were not aware of it).
None Evaluation found no evidence that change had occurred	**None** Evaluation found no evidence of NRC influence
Low NRC advocacy 'ask' considered, possibly planned, but little change occurred to date.	**Low** NRC Influence was just one of many possible influences on target.
Medium NRC advocacy 'ask' considered and some implementation seen (e.g. pilots), but not yet widespread or systematic.	**Medium** Influence of NRC was one of a limited number of possible influences on target
High NRC advocacy 'ask' considered and integrated; potential for sustainable and long-term change.	**High** NRC was the key or only influence on target

Source: O'Neil, and Goldschmid, 2014:27

The evaluation also used the Theory of Change that was developed to inform NRC's programme in the area of access advocacy and 'checked' the DRC's evaluation results against it to gauge which piece of evaluative evidence could be situated against the ToC to support (or not) the expected linkages between different levels of results (pp.11-12).

The evaluation offers an example of the benefit of syncing evaluation timing with other broader policy development processes. In this case, the evaluation of NRC's advocacy work and protection in DRC was timed with the drafting process of the the NRC's Global Protection Policy.[8]

One of the challenges was the ambition of tracking 'too many' changes (30 in total) at 'too many' levels. In the end, the evaluation team recommended reducing the number of expected changes to ten in order to facilitate the tracking of contribution and allow for deeper analysis of fewer most significant changes.

IV. Concluding reflection on so-called 'gold' and 'platinum standards'

As Jane Davidson, put it 'yes' there is a so-called 'gold standard'. And it should not be about specific choice of methods and preference for certain approaches to evaluation. The gold standard should be about systematically pursuing causal reasoning in evaluation (Davidson, 2005, 2009).

Put succinctly, this means that the evaluation questions should drive method and design choices. Not the other way around.

In line with those views, Patton (2014a; b) has recently shared some reflections on the questions around the quality to aim for when selecting and applying different approaches and designs to answer causal questions in evaluation. He argued that 'we need to (…) aim to supplant the gold standard with a new platinum standard: methodological pluralism and appropriateness.' (Patton, 2014 b).

Davidson (2000, 2005) suggest asking a key question that needs to be answered in order to decide which design, method or analytical tool offer the best fit to be used in an evaluation. We need to ask:

8 ALNAP's study on 'Using evaluation for a change' (Hallam and Bonino, 2013) identified evaluation timing issues and ensuring the evaluative work is connected with other organisational processes – such as policy development or policy revisions – as one of the main ingredients that promote evaluation use.

Which level of certainty the evaluation commissioning agency and the intended users of the evaluation need to have in order to be confident that a given intervention led to a certain change or result?

"The ultimate goal – or platinum standard, to use Patton's quote– is to identify and propose the most appropriate blend or best fit of designs and methods to answer the evaluation questions at hand."

In many organisational realities, decision-makers are prepared to make decisions if they were, say, 70% or 80% certain of the evidence provided to prove or disprove a given evaluative statement. Different contexts and different types of decision will call for different 'threshold' of certainty. And indeed, because each decision context requires a different level of certainty, it is important to be clear up front about the level of certainty required by decision-makers and other evaluation stakeholders (Davidson, 2005: 69).

The depth and breadth of the required evidence base is a key consideration in evaluation planning and should be based on a thorough assessment by the evaluator of stakeholder information needs. This not only will help with budgeting the evaluation more accurately but also will facilitate any up-front discussions with the client about the trade-offs between budgets, time lines, and the certainty of conclusions (Davidson, 2000: 25).

Having some clarity on this type of stakeholders' expectations should thus help evaluators better explain – and sometime defend – their design and method choices. It should also help evaluators show the trade-offs between, budget allocated to the evaluation, evaluation time-lines, skills, and resource required by different types of approaches. The ultimate goal – or platinum standard, to use Patton's quote– is to identify and propose the most appropriate blend or best fit of designs and methods to answer the evaluation questions at hand.

References

Befani, B. (2012) Models of causality and causal inference. in Stern et. al (2012) London: Department for International Development.

Cartwright, N. (2007) Hunting causes and using them: Approaches in philosophy and economics. Cambridge, UK: University Press.

Chambers, R. et al. (2009) Designing impact evaluations: different perspectives. 3ie Working Paper no.4. New Delhi: International Initiative for Impact Evaluation.

Cook, T. (2000). The false choice between theory-based evaluation and experimentation. New Directions for Evaluation, Vol. 87 pp. 27-34

Cook, T. et al. (2010) Contemporary thinking about causation in evaluation: A dialogue with Tom Cook and Michael Scriven. American Journal of Evaluation, 31(1), 105-117.

Coryn, C. (2013) Experimental and quasi-experimental designs for evaluation. Washington DC: American Evaluation Association. Training materials.

Davidson, J. (2000) 'Ascertaining causality in theory-based evaluation' in New Directions for Evaluation - Special Issue: Program Theory in Evaluation: Challenges and Opportunities. Vol. 2000 (87) pp. 17–26. http://onlinelibrary.wiley.com/doi/10.1002/ev.1178/abstract

Davidson, J. (2005) Evaluation Methodology Basics: The Nuts and Bolts of Sound Evaluation. Thousand Oaks: Sage

Davidson, J. (2009) Causal inference: Nuts and bolts, Presentation for ANZEA evaluation conference. http://realevaluation.com/pres/causation-anzea09.pdf

Davidson, J. (2013) Understand Causes of Outcomes and Impacts. American Evaluation Association (AEA) Coffee Break Demonstration CBD141 delivered on 21 March 2013. http://comm.eval.org/Resources/ViewDocument/?DocumentKey=a2b20160-c052-499d-bdb5-0ae578477d2a

EES (European Evaluation Society) (2008) ESS Statement: The Importance of a Methodologically Diverse Approach to Impact Evaluation – Specifically with Respect to Development Aid and Development Interventions. EES, December 2008.#

Gerring, John (2012) Social Science Methodology: A Unified Framework: Second edition (Cambridge, UK: Cambridge University Press).

Goertz, G. and Mahoney, J. (2012) A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences. Princeton University Press. http://press.princeton.edu/titles/9898.html

Hughes, K. and Hutchings, C. (2011) Can we obtain the required rigour without randomisation? Oxfam GB's non-experimental Global Performance Framework. 3ie Working Paper no. 13. New Delhi: International Initiative for Impact Evaluation (3ie)

Jones, H (2009) The 'gold standard' is not a silver bullet for evaluation Opinion Piece. London: ODI.

Karlan, D. (2009) 'Thoughts on Randomized Trials for Evaluation of Development: Presentation to the Cairo Evaluation Clinic', in Chambers, et al. (2009) Designing impact evaluations: different perspectives. 3ie Working Paper no.4. New Delhi: International Initiative for Impact Evaluation. pp. 8-13.

Knox Clarke, P. and Darcy, J. (2014) Insufficient evidence? The quality and use of evidence in humanitarian action. ALNAP Study. London: ALNAP/ODI.

Lehmann, C. and Masterson, D. (2014) Emergency Economies: The Impact of Cash Assistance in Lebanon. An Impact Evaluation of the 2013-2014 Winter Cash Assistance Program for Syrian Refugees in Lebanon. Beirut, Lebanon: International Rescue Committee.

Mathison, S. (ed.) (2005) Encyclopedia of Evaluation. London and New York: Sage. http://srmo.sagepub.com/view/encyclopedia-of-evaluation/n398.xml

Mayne, J. (2012). Making causal claims, ILAC Brief No. 26. Rome: Institutional Learning and Change (ILAC) Initiative. http://www.cgiar-ilac.org/files/publications/mayne_making_causal_claims_ilac_brief_26.pdf

O'Neil, G. and Goldschmid, P. (2014) Final Report Evaluation of NRC's 2012-13 protection and advocacy work in the DRC. Oslo: Norwegian Refugee Council http://www.nrc.no/default.aspx?did=9182709#.VFeh0RZvnKc

Patton, M. Q. (2012) Contextual Pragmatics of Valuing, New Directions for Evaluations, Vol 2012 (133), 1-129.

Patton, M. Q. (2014a) Qualitative Research & Evaluation Methods - Integrating Theory and Practice (Fourth Edition). Thousand Oaks: Sage Publ. http://www.sagepub.com/books/Book232962#tabview=toc

Patton, M.Q. (2014b) Week 47: Rumination #3: Fools' gold: the widely touted methodological "gold standard" is neither golden nor a standard, posted on BetterEvaluation on 4th December 2014 at: http://betterevaluation.org/blog/fools_gold_widely_touted_methodological_gold_standard

Proudlock, K. Ramalingam, B. with Sandison, P. (2009) 'Improving humanitarian impact assessment: bridging theory and practice', in ALNAP 8th Review of Humanitarian Action. London: ALNAP/ODI (www.alnap.org/ resource/5663.aspx).

Puri, J, Aladysheva, A, Iversen, V, Ghorpade, Y and Brück, T, 2014. What methods may be used in impact evaluations of humanitarian assistance?, 3ie Working Paper 22. New Delhi: International Initiative for Impact Evaluation (3ie)

Reichardt, S. (2005) 'Experimental design' in Mathison, S. (ed.) Encyclopedia of Evaluation. London and New York: Sage. http://srmo.sagepub.com/view/encyclopedia-of-evaluation/n398.xml

Rogers, P. (2000) Program theory: Not whether programs work but how they work. In D. L. Stufflebeam, G. F. Madaus, & Kellaghan, T, (Eds.) Evaluation models viewpoints on educations and human services evaluation 2nd ed. (209-233). Boston, MA: Kluwer Academic Publishers.

Rogers, P. (2009) Matching Impact Evaluation Design to the Nature of the Intervention and the Purpose of the Evaluation, Journal of Development Effectiveness, Vol 1 (3), 217-226.

Scriven, M. (2010) 'The Calorie Count of Evaluation', in Journal of MultiDisciplinary Evaluation, Vol. 6(14) p. i-v, <http://journals.sfu.ca/jmde/index.php/jmde_1/article/view/287/290>.

Scriven, M. et al. (2010) Contemporary thinking about causation in evaluation: A dialogue with Tom Cook and Michael Scriven. Cited in Cook, T. et al. 2000 in American Journal of Evaluation, Vol. 31(1), 105-117.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002) Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton-Mifflin Company.

Stern, E. (2008) Current Thinking about Impact Assessment. Presentation delivered at the ALNAP 24th Biannual meeting. Available at: http://www.alnap.org/meetings/24.htm

UN OIOS-IED (2014) Inspection and Evaluation Manual. New York: United Nations Office of Internal Oversight Services (OIOS), Inspection and Evaluation Division (IED).

WFP and UNHCR (2012) Synthesis Report of the Joint WFP and UNHCR Impact Evaluations on the Contribution of Food Assistance to Durable Solutions in Protracted Refugee Situations. Rome and Geneva: WFP-Evaluation Office and UNHCR/PDES. http://www.wfp.org/node/383882

White, H. and Phillips, D.(2012) Addressing attribution of cause and effect in small n impact evaluations: towards an integrated framework. 3ie Working Paper no. 16. New Delhi: International Initiative for Impact Evaluation (3ie)

This Method Note is the result of a discussion on 'Ensuring quality of evidence generated through participatory evaluation in humanitarian contexts', which took place among members of the ALNAP Humanitarian Evaluation Community of Practice (CoP) between May and September 2014.

The CoP discussion presented in this brief note has been enriched with references to evaluation literature and guidance provided by Jessica Alexander and Francesca Bonino. Whenever possible, these notes highlight the experience from ALNAP Members in dealing with specific evidential challenges in humanitarian evaluation practice.

ALNAP's Humanitarian Evaluation CoP provides a space to:

- Discuss issues related to humanitarian evaluation
- Share resources and events
- Consult peers

YOU CAN JOIN THE CoP HERE:

https://partnerplatform.org/humanitarian-evaluation