ALNAP

Evaluating protection in humanitarian action:

Decision-making processes, common issues and challenges

Ian Christoplos and Francesca Bonino

ALNAP PILOT GUIDE

ALNAP is a unique system-wide network dedicated to improving the performance of humanitarian action through shared learning. **www.alnap.org**

Suggested citation
I. Christoplos & Bonino, F. (2016) Evaluating protection in humanitarian action: Focus on decision-making processes and options to address common issues and challenges. ALNAP Pilot Guide. London: ALNAP/ODI.

ISBN 978-1-910454-44-2

Irish Aid
Department of Foreign Affairs
An Roinn Gnóthai Eachtracha

About the authors

Francesca Bonino is a former Research Fellow with the ALNAP Secretariat based in London.

Ian Christoplos is a professional evaluator working with the consulting firms NIRAS-Indevelop and Glemminge Development Research, and a researcher with the Danish Institute for International Studies.

Acknowledgements

Contributions: Rachel Goldwyn

ALNAP internal review: Paul Knox Clarke, Luz Saavedra, Alexandra Warner

Research support: Sian Cook

Experts consulted and interviewed during the development of first draft of this guide: Jon Bennett, Margie Buchanan-Smith, Vanessa Corlazzoli, John Cosgrave, Jamie Khalil, Mark Rogers, Seb Taylor.

External reviewers: John Cosgrave and Margie Buchanan-Smith, Jock Baker (independent), Eva Svoboda (HPG), Isabella Jean who shared the guide with some of her colleagues working on Do no Harm issues (CDA), Velina Mikova (ICRC), Mikkel Nedeegard and Kathrine Starup (DRC), Brennan Weber (DRC Lebanon office), Gita Swamy Meier-Ewert (Senior Planning and Monitoring Specialist - UNHCR), Arne Treves (UNCHR / GPC), Rachel Hastie (Oxfam GB), Erica Mattellone (UNICEF), Jessica Lenz (InterAction), Katrien Ringelé and Gergey Pasztor (IRC), Siri Elverland and Lian Bradley (NRC), Langdon Greenhalgh and Caroline Andersen (USAID-OFDA).

Publications and communications: Alex Glynn
Copyediting: Aaron Griffiths
Design and typesetting: Chloé Sanguinetti

Contents

List of boxes, figures and tables

Acronyms

ALNAP	Active Learning Network for Accountability and Performance in Humanitarian Action
CA	Contribution Analysis
DE	Development Evaluation
DFID	Department for International Development
EHA	Evaluation of Humanitarian Action
FGD	Focus group discussion
GBV	Gender-based violence
MSC	Most Significant Change
OM	Outcome Mapping
PIA	Participatory Impact Assessment
QCA	Qualitative Comparative Analysis
RDD	Regression Discontinuity Designs
RE	Realist Evaluation
RTE	Real-Time Evaluation
SCM	Success Case Method
SNA	Social Network Analysis
TBE	Theory-Based Approaches to Evaluation
U-FE	Utilisation-Focused Evaluation

INTRODUCTION

SECTION 1: Why an evaluating humanitarian action companion guide on protection?

> How can people and communities at risk in situation of crisis and conflict be better protected? And what role can humanitarian agencies and their staff play in helping bringing about such protection on the ground? (Slim and Bonwick, 2005: 11)

Despite the stated centrality of protection in humanitarian action (IASC, 2013) and a growing attention to protection activities, the evaluation of protection has received relatively little attention. This pilot guide seeks to fill this gap, providing insights and guidance to those evaluating protection in the context of humanitarian action.

Evaluating protection in the context of humanitarian action[1] can be seen as a nascent but growing subset of practice within the broader fields of Evaluation of Humanitarian Action (EHA)[2] and the evaluation of protection outside of humanitarian action.

The scoping paper that led to the development of this pilot guide (Bonino, 2014) highlighted that compared to aspects of EHA practice, and for which more evaluation-specific guidance is available, there is a dearth of understanding and guidance in relation to evaluating protection in humanitarian action.

For example, a brief review of inter-agency guidance materials on humanitarian evaluation (Bonino, 2014: 38-41) uncovered the rather different and varying ways in which protection in humanitarian action has been dealt with in humanitarian evaluation guidance as:

- One of the cross-cutting issues that humanitarian evaluations can examine (IASC, 2011)
- As a programme theme with initiatives that are solely focused on protection
- An overarching theme that all humanitarian evaluators should consider to some extent in their EHA work – regardless whether it is explicitly requested in the evaluation terms of reference (Hallam, 1998).

Evaluation guidance focusing on protection in humanitarian action is limited, fragmented and confined to agency-, sector- and theme-specific programming manuals that often give limited overall guidance on the specific challenges of looking at protection.

The practice of evaluating protection (either as a primary or secondary line of evaluation inquiry) appears rather dispersed and inconsistent[3].

1 The objectives of humanitarian action are to save lives, alleviate suffering and maintain human dignity during and in the aftermath of human-induced crises and natural disasters, as well as to prevent and strengthen preparedness for the occurrence of such situations. (Cosgrave and Buchanan-Smith, 2016).

2 Evaluation of humanitarian action (EHA) is defined as the systematic and objective examination of humanitarian action, to determine the worth or significance of an activity, policy or programme, intended to draw lessons to improve policy and practice and enhance accountability. (Cosgrave and Buchanan-Smith, 2016).

3 This point was also highlighted in a useful scoping paper (Reichhold, Binder and Niland, 2013) that searched the ALNAP evaluation library (ALNAP Humanitarian Evaluation and Learning Portal – www.alnap.org/resources) to map and review the evaluations that look at protection.

This pilot guide, a companion to the ALNAP EHA guide (Cosgrave and Buchanan-Smith, 2016), offers evaluation-specific insights that can speak to the diverse membership within the ALNAP network who work within and around protection, both as specialists and 'generalists'. It does not attempt to define protection but is rather intended as support for evaluators and evaluation managers involved in analysing interventions that take their points of departure from a variety of definitions. It can be noted that the Global Protection Cluster describes this broad scope of protection as consisting of:

> an objective, a legal responsibility and a multi-sector activity to (1) prevent or stop violations of rights, (2) ensure a remedy to violations- including the delivery of life-saving goods and services- and (3) promote respect for rights and the rule of law. (GPC, 2015)

About the pilot process of the ALNAP Guide on evaluating protection in humanitarian action

Over 2016 and into 2017, the guide will be piloted to further its utilisation-focus; specifically, the aim of the pilot is to identify any thematic gaps, additional resources or tools, and capture further practical examples.

For more information on the pilot process for this guide, please visit http://www.alnap.org/evaluating-protection or email the ALNAP Secretariat at eha@alnap.org.

Objectives

The scoping phase for this pilot guide (Bonino, 2014) found that theory and practice to date do not adequately articulate how EHA could become more responsive to the specific features of protection in humanitarian action that make evaluation challenging.

This pilot guide provides evaluation teams and staff in evaluation commissioning positions and in evaluation management and advisory roles (the primary audience of this guide) with tools to navigate those challenges. More specifically, the objectives are to:

- Focus on the **decision-making processes** around evaluating protection in humanitarian action and presenting the **critical decision points** in an evaluation where the focus includes protection.
- Describe some of the **trade-offs required and options available** to evaluators and evaluation commissioning offices in preparing for an evaluation, selecting approaches and methods and gathering data.
- Offer **practical insights, tools and approaches** that can be used in evaluating protection in humanitarian action.

SECTION 2: Main features of this pilot guide

Target users: the pilot guide addresses a broad audience of evaluators, staff in evaluation commissioning positions and staff in evaluation management and advisory roles (in agencies with and without a specific protection mandate).

Agency staff in protection programming advisory and support roles who are not 'evaluation experts' are often called upon to support the management of evaluations, comment on evaluation terms of reference, review evaluation proposals, be part of mixed evaluation teams, and provide technical support to evaluations that are commissioned specifically to look at protection programming in a given humanitarian context. Some familiarity with key terms and concepts in EHA and in protection is assumed. These key terms are presented in "SECTION 3: Brief orientation on protection in humanitarian action" on page 13.

The guide can be read from start to finish, but we suggest users zoom in on different modules depending of the relevant stage of the evaluation process:

MODULE A	MODULE B	MODULE C
Initiating the evaluation of protection in humanitarian action	Data collection in the evaluation of protection in humanitarian action: Practical and ethical considerations	Analysis

This guide has been designed to:

- Focus specifically on decision-making processes and options in an evaluation process.
- Speak to evaluators, evaluation managers, and staff in evaluation-commissioning roles working in both 'protection specialist' and 'generalist' positions within agencies with or without a specific protection mandate in different operational environments, as well as to independent evaluators.

This guide has not been designed to:

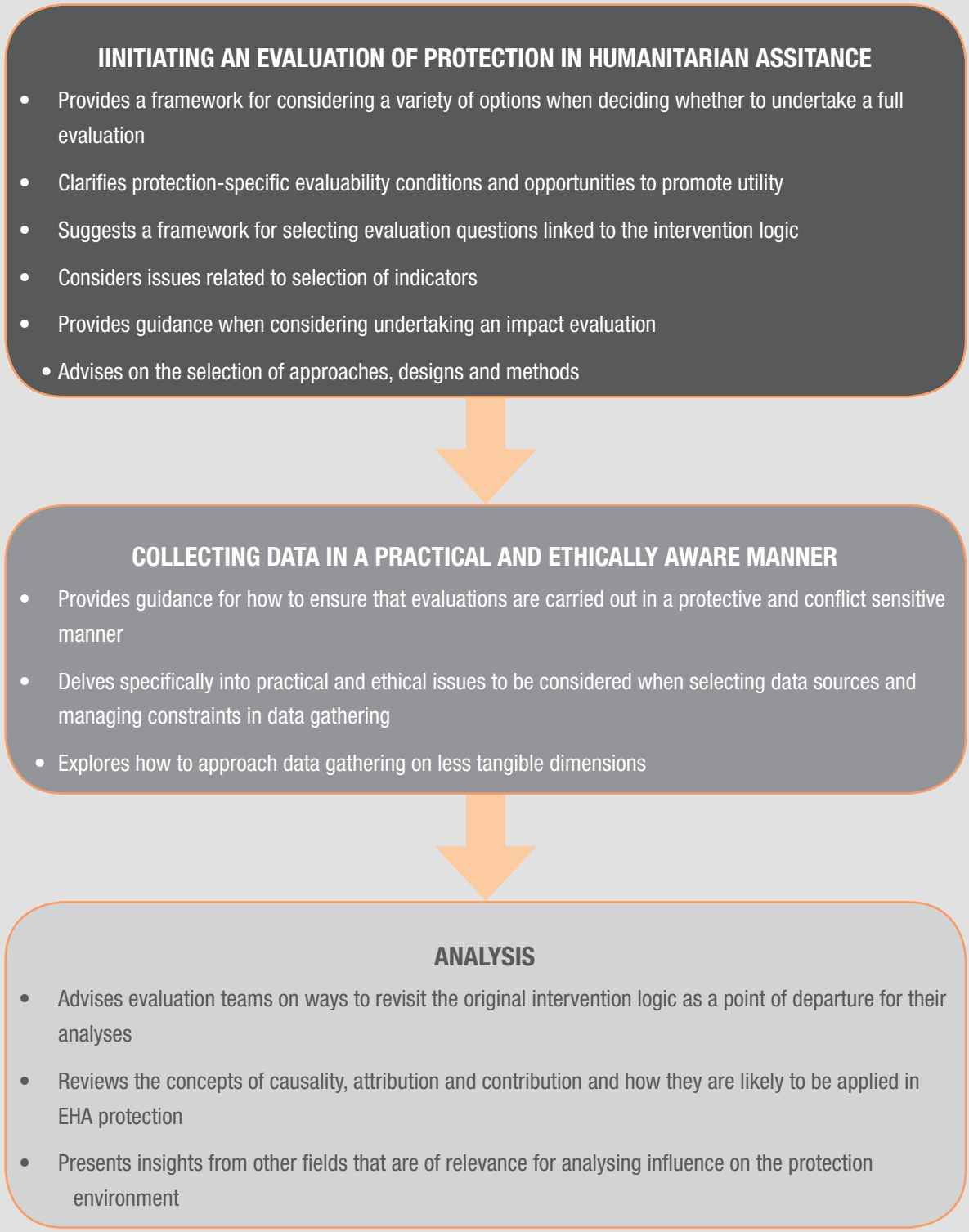
- Be a complete programming guide covering all the steps in a programming cycle. It focuses on specific decisions relating to initiating and scoping, designing and undertaking evaluations that look at protection.
- Duplicate the content covered in the ALNAP EHA guide, which remains the main entry point to and reference text on humanitarian evaluation for ALNAP.

Box 1: Features of the evaluation of protection in humanitarian action pilot guide

To improve accessibility and help navigation of its content, this guide features:

- A **detailed content map (**See Figure 1**)** presented as a flowchart that features all the components of the guide and indicate to which stage of a generic evaluation (and pre-evaluation) process they refer to.
- A number of boxes featuring **evaluator insights**: These are short nuggets from evaluation practitioners reflecting on the use they made of a specific tool or framework presented in the guide.
- A **toolkit section** to describe in more detail selected tools and approaches mentioned in the main body of the guide. Note that further tools are available at www.globalprotectioncluster.org/en/tools-and-guidance/essential-protection-guidance-and-tools.html.
- Whenever a specific tool or section is mentioned in the main body text of the guide, it is indicated with a red highlight.

Figure 1: Evaluating humanitarian protection guide flowchart



SECTION 3: Brief orientation on protection in humanitarian action

While the Inter-Agency Standing Committee (IASC) Principals are committed to 'ensuring the centrality of protection in humanitarian action' (IASC, 2013), diverse agency mandates and/or priorities around protection in humanitarian action have significant implications for the overall scope of the evaluation and the selection of analytical frameworks.

The **concept of 'protection' is itself multifaceted** and defies clear categorisation and linear measurement.

The different aspects of protection in humanitarian action shape a given agency's **institutional and policy orientations towards protection in a given intervention**. This has practical repercussions for what gets analysed and monitored and the processes and results that are evaluated.

Acknowledging these issues, the International Committee of the Red Cross proposed to use the so-called **egg framework on protection**[4] (Figure 2) as a way of showing the relations among the different strands of protection work in humanitarian contexts. This framework specifies three main families of protection actions in **humanitarian action** (Giossi Caverzasio, 2001: 21-24). The relations among the different actions are conceived as being inter-dependent, but non-hierarchical. It is also expected that different actions may be carried out simultaneously[5]. Evaluators may be asked to look at activities, services and expected results in all three of them:

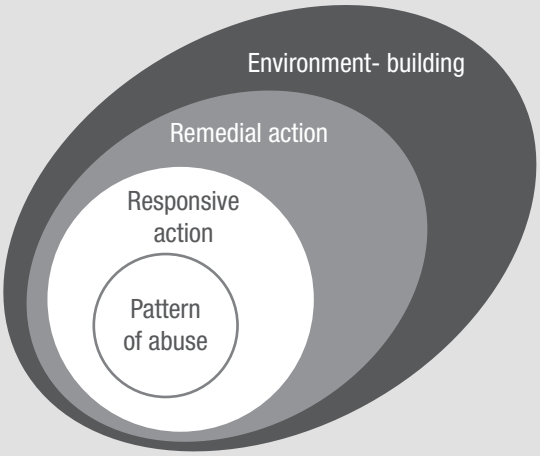
- **Responsive actions** to stop, prevent the recurrence of, or alleviate the immediate effects of an emerging or established pattern of abuse.
- **Remedial actions** being undertaken after abuse has occurred to restore people's dignity and ensure adequate living conditions.
- **Environment-building actions** to foster a political, social, cultural, institutional, and legislative environment that enables or encourages national authorities to fulfil their obligations and respect individual rights.

As more attention is paid to strengthening sustainable national and local institutional capacities to undertake protection actions (beyond humanitarian action), **environment-building** has come to encompass an expanding range of support that recognises and seeks to reinforce the often under-acknowledged role of communities and address the central role of the state (either positive or negative) in protection.

4 Although the egg framework emerged from an exchange among agencies with protection mandates, it is now widely used by non-mandated agencies as well (e.g. Allaire, 2013).

5 Preventive actions are also considered, even if different humanitarian actors see their role in different lights in this area.

Figure 2: The egg protection framework



Source: Giossi Caverzasio (2001: 21).

The contexts in which protection actions are carried out have major implications for selection of evaluation indicators that unpack the relevance and define the intended results of a given approach to protection. Major factors include:

- The willingness and capacity of the state and the authorities to respond to protection risks and violations
- The capacity of civilian communities to help themselves and their space to act
- The agency's capacity to respond
- The risk the action would create for the civilian population's security
- The political risk it would create for the agency's security and access
- The duration of the action
- The agency's experience with similar actions in a given setting
- The activities and mandates of other actors.

A study commissioned by the Global Protection Cluster neatly captures the multifaceted nature of protection in the context of humanitarian action[6]:

6 Arguably, this is one of the reasons why developing a taxonomy measurement and programming around protection in HA, as well as a related taxonomy to guide humanitarian evaluators looking at protection continues to prove challenging. The independent whole-of-system review of protection offers this reflection: 'Ambiguity surrounding the essence of effective protection programming can give rise to unhelpful illusions that anything and everything can be deemed to be protective. The all-encompassing nature of the formal definition fuels confusion. The absence of a common understanding or agreed operational approach to protection in the context of humanitarian action works against sound needs assessments, strategic prioritisation, coordination and the ability to monitor and evaluate programme implementation including outcomes.' (Niland et al., 2015: 23)

> » Protection defies neat labelling because it is at the same time the goal underlying the whole humanitarian response (the reason for humanitarian action), an approach or lens on the humanitarian response (a way of understanding all dimensions of humanitarian endeavour), and a more narrowly-defined family of activities that aim to prevent and mitigate threats to vulnerable persons. (Murray and Landry, 2013: 4; emphasis in the original text)

This complicates the evaluator's tasks of identifying indicators and tailoring methods to appropriately assess and judge intended results. It also creates challenges in delimiting and describing what the evaluation needs to look at in several ways.

Firstly, when asked to 'evaluate protection', a specific set of protection results, or a protection component of a larger programme or intervention, it is essential to clarify the types of protection included in the intervention, including how the concept of protection is used by the agency. Where different areas of protection and perceived priorities are combined in a given intervention, it is important for evaluators to revisit how these have been delineated.

Secondly, depending on the purpose of the evaluation, the questions it asks and the orientation of who commissioned it, evaluators may be asked to look at the scope of protection in humanitarian action as:

- An overarching theme of analysis for a whole response in a given humanitarian crisis or conflict context
- A specific issue that cuts across different (sectoral) programming areas and interventions
- A primary line of inquiry in an evaluation looking at sector-specific results in a dedicated area of programming (e.g. around child protection, gender based violence and protection against sexual exploitation and abuse)
- A secondary line of inquiry in an evaluation that looks at relevance and quality dimensions of a given response or programme.

Thirdly, there can be evaluation scenarios where terms of reference do not actually mention 'protection' per se, but where protection is nonetheless an implicit focus. (Note that, according to recent IASC statements, protection is 'central' to humanitarian action whether explicit or not.) In those cases, the evaluation team may need to tease out the protective features in a programme that can be inferred from, for example, 'do no harm' measures or the safety and accessibility of the service or assistance provided.

MODULE A – Initiating the evaluation of protection in humanitarian action

» This module is addressed to those commissioning or planning an evaluation, and evaluation teams during inception phases.

Content of this module at a glance

	What is this section on?	Who is this section for?
SECTION 4:	Provides guidance for considering a spectrum of reflective and evaluative options when deciding whether to undertake a full evaluation	Primarily for evaluation offices /evaluation commissioning staff
SECTION 5:	Clarifies the protection-specific evaluability conditions and opportunities to promote utility	For both evaluators and evaluation offices
SECTION 6:	Suggests a framework for selecting evaluation questions linked to the intervention logic	Primarily for evaluation office and commissioning staff, but also useful for evaluators during inception phases
SECTION 7:	Suggests a framework for selecting evaluation questions linked to the intervention logic	Primarily for evaluation office and commissioning staff, but also useful for evaluators during inception phases
SECTION 8:	Presents issues to be reviewed when considering an impact evaluation	Primarily for evaluation office and commissioning staff, but also useful for evaluators during inception phases
SECTION 9:	Provides guidance in the selection of approaches, designs and methods	Primarily for evaluation teams, but also for offices /evaluation commissioning staff when drafting terms of reference and assessing inception reports

'Initiating' an evaluation here refers to the different actions that could be usefully considered *before an evaluation* and *during its inception phase*. The proportion of tasks undertaken before the evaluation and during the inception phase is likely to vary in different organisations and assignments.

Investing in pre-evaluation and inception processes has emerged as an area of good practice[7] that can encourage evaluation utility by being:

- Better understood and more easily accepted and ideally 'owned' by its primary intended users
- More useful to its ultimate users
- Better supported by programme staff and championed by the leaders and managers who should take action on the evaluations' conclusions and recommendations.

7 See for instance, see Hallam and Bonino (2013) for a study specific to humanitarian evaluation practice, and Rist, Boily, and Martin (2011) and Heider (2011) for some useful insights from broader development aid evaluation practice.

This first module should be seen as a reminder that the path to improving the overall quality, usefulness and credibility of EHA-protection can start with:

» **Considering a spectrum of reflecting and evaluative options** and considering the best fit with purpose and questions that stakeholders have about protection (discussed in "SECTION 1: Why an evaluating humanitarian action companion guide on protection?" on page 8), focusing on the intended uses and users.

» **Ensuring that evaluators are critically reflecting** on the systemic and organisation-specific features of protection in the context of humanitarian action.

» **Clarifying the protection-specific evaluability conditions** to take more informed and better timed decisions around initiating and scoping an evaluation or reflective exercise.

» **Giving space for the evaluation team** to use the inception phase to build consensus around evaluation objectives and focus, thereby reinforcing ownership and opportunities for maximising utility.

» **Ensuring that approach, design and methods** are suited to the evaluation questions, expectations and field conditions facing the evaluation team.

SECTION 4: The spectrum of reflective and evaluative options

In addition to formal, fully fledged evaluations that aspire to systematically abide by evaluation standards (UNEG, 2005) and follow process and product quality assurance processes (Yarbrough, et al., 2011), there are other types of exercises that 'promote active reflection infused with evaluative thinking' (Scharbatke-Church, 2011b: 7).

» Evaluative thinking can underpin a wide **spectrum of evaluative and other reflective processes**, not just formal evaluations.

The 'illustrative spectrum'[8] is visualised in Figure 3. The menu of options presented is broad brush and not meant to replace other typologies or different approaches to classifying EHA[9].

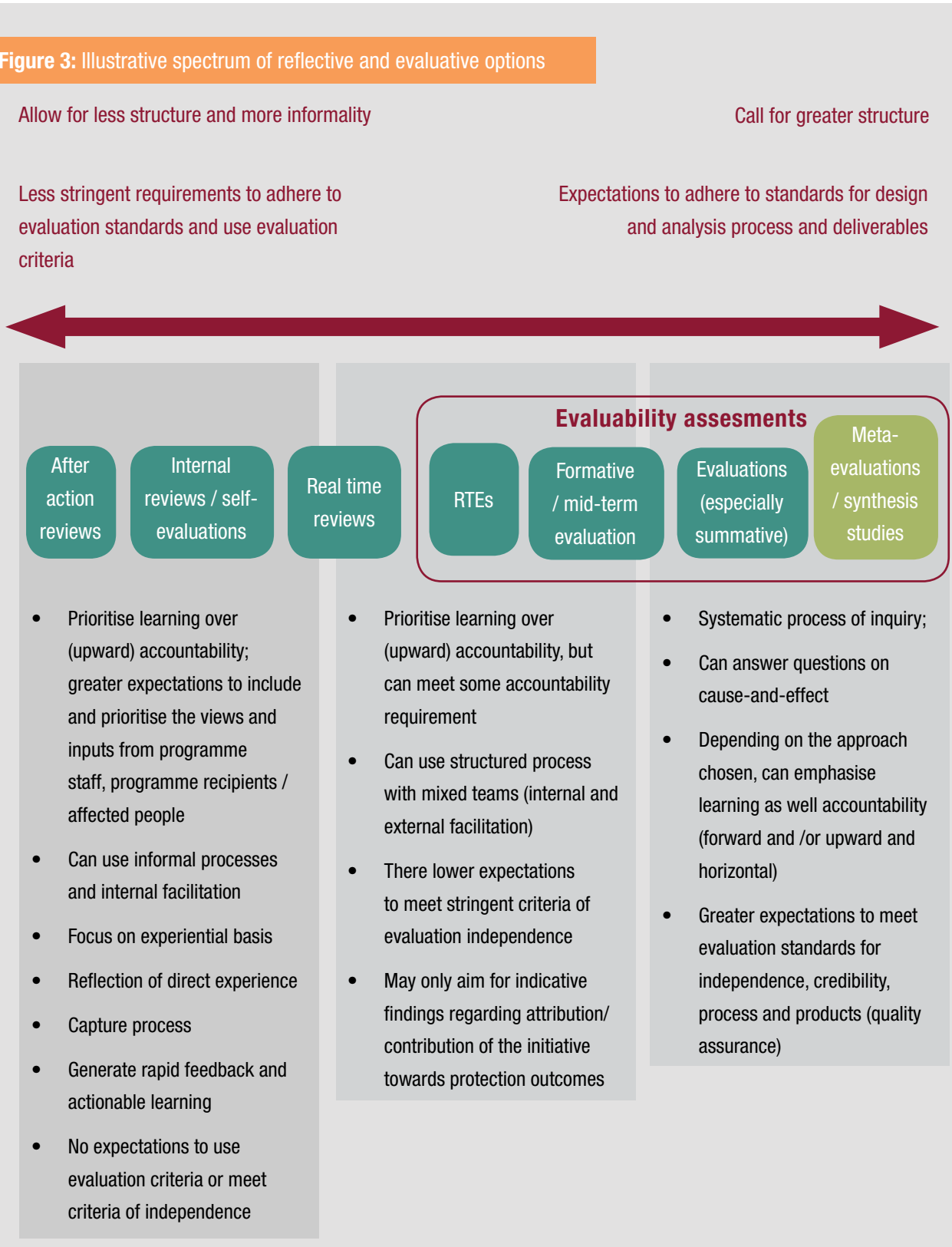
The spectrum moves left to right from informal, experience-based, after-action reviews towards more structured evaluations that systematically apply quality standards and criteria.

It is important to clarify that by presenting this menu of options and highlighting possible alternatives to fully fledged evaluations (especially of the summative kind) is not intended to suggest shortcut solutions to 'replace' evaluation with second-best options.

8 This is what Scharbatke-Church first called it in her writing on peacebuilding evaluation (Scharbatke-Church, 2011)

9 Programme evaluation literature abounds with more of less detailed typology of evaluation based on purpose, dominant design orientation, approaches used, expected uses and users. This guide suggest one way of looking at typology in evaluation, acknowledging that in EHA, different agencies often use slightly different terminology and approaches to classification. For an example see IFRC, 2011a.

Figure 3: Illustrative spectrum of reflective and evaluative options



NOTES: This spectrum does not cover the relative accountability and learning gradients in different exercises. It does not touch on expectations around quality and robustness of evidence generated through the various types of exercises. Also note that under the label of 'evaluation' other specific typologies are nested along different characteristics depending on evaluation scope, governance arrangement, thematic and programmatic focus, type of questions asked, timing etc.

Different types of evaluative exercises serve different purposes. They answer different types of questions and serve different stakeholders and expected users.

When taking a decision about which option along the spectrum should be considered for evaluating protection in humanitarian action, we should ask which option offers the best fit in terms of:

a. Overarching evaluation purpose(s)
b. Type and balance of questions asked
c. Evaluability conditions
d. Resources available
e. Timing and stage in the programme cycle, and stage in the humanitarian response cycle in which the intervention is situated
f. 'Maturity' of the intervention (e.g. innovation project, pilot stage, consolidation, scale up, scale down, exit)
g. Stakeholder demands and expectations in terms of evidence generated through the evaluative or more reflective exercise
h. Stakeholder demands and expectations in terms of the learning and accountability (upward towards funders, horizontal towards partners, forward towards affected population) generated through the evaluation.

SECTION 5: Protection-specific evaluability conditions and opportunities to promote utility

Why evaluability and utility analysis matters

Clarifying the protection-specific evaluability conditions in a programme or intervention in pre-planning processes and inception phases is critical in EHA-protection to help orienting the decisions around:

- Institutional and programme readiness for the evaluation, including in terms of timing.
- The feasibility of including protection-specific questions (as the primary or secondary line of inquiry in an evaluation) at a given time in the life of the programme.
- How to maximise the value and usefulness of an evaluation by focusing on intended use by intended users.

This pilot guide emphasises the role that the **analysis of evaluability and utility** can play in planning and initiating evaluation of protection in humanitarian action to pave the way to better designed, better-timed and scoped evaluations[10].

An analysis of evaluability may also help to uncover (and potentially tackle) any friction between the evaluation's requirements for upwards accountability and other expectations around learning and use.

Defining evaluability analysis for EHA-protection:

Evaluability analysis can be defined as a structured process of description and synthesis upon which to recommend whether an evaluation (of any type) is feasible and whether conditions are present to answer the questions the commissioners of the evaluation are asking.

Most commonly, evaluability analysis involves asking a series of questions (usually through a set of checklists) about the main factors and variables that are likely to influence timing, feasibility, institutional readiness and level of ambition attached to an evaluation of protection in humanitarian action.

Evaluability analysis usually covers:

- The level of ambition and type of questions that evaluation stakeholders and programme stakeholders would like to see answered
- Intervention logic / programme logic
- The availability and quality of information already generated by the programme
- Conduciveness of context for evaluating protection in humanitarian action (including programme and institutional readiness to the evaluation)
- Expectations regarding the evaluation among different users and stakeholders.

10 The evaluability considerations presented here are arguably also relevant to general EHA practice.

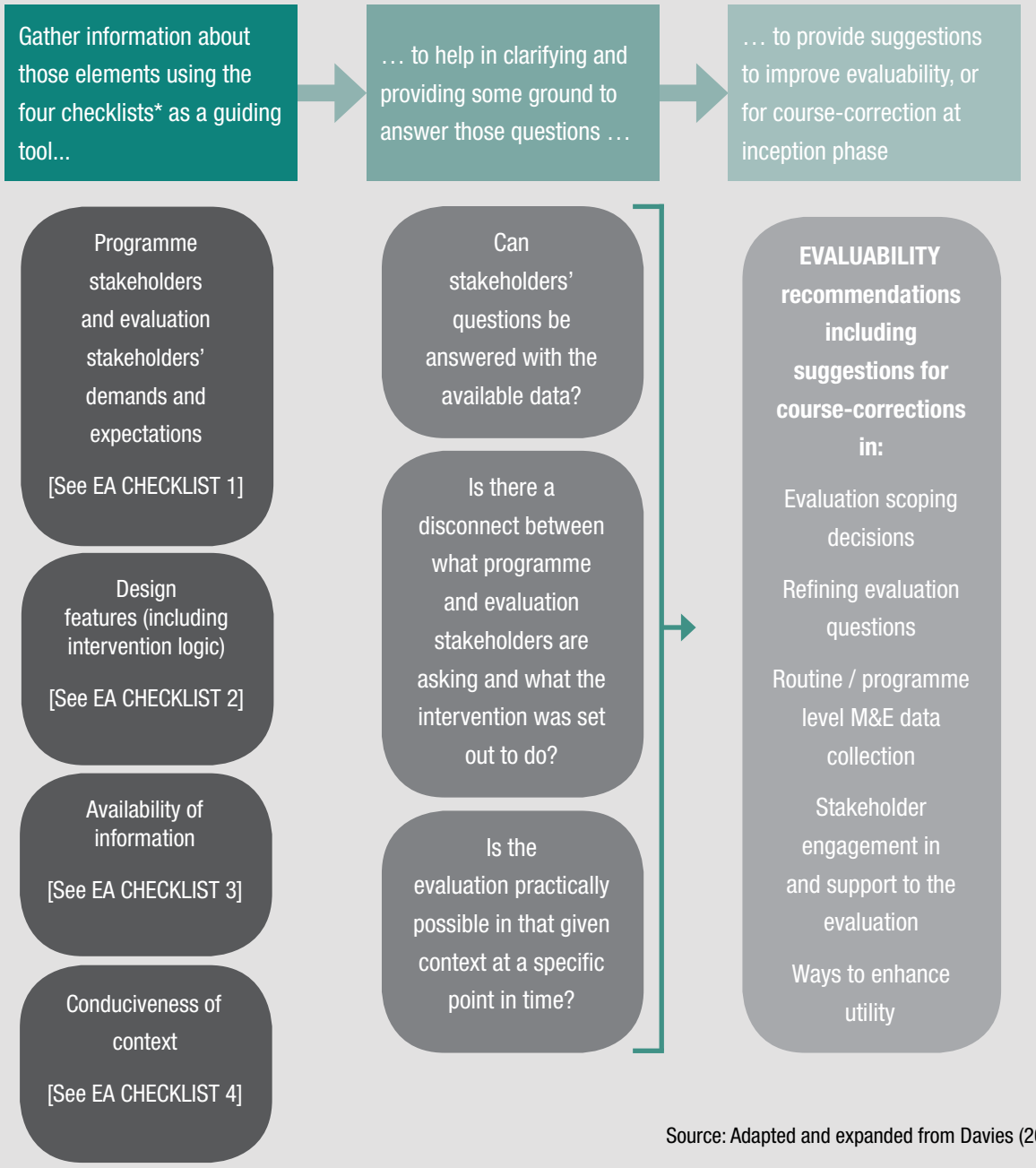
Main aspects to consider to judge evaluability and enhance utility

Four aspects should be assessed in analysing evaluability in evaluating protection in humanitarian action[11]: See "TOOLKIT item #1 – Evaluability checklists for evaluation of protection in humanitarian action" on page 77

- **The overall level of ambition** and issues that evaluation stakeholders and programme stakeholders would like the evaluation to explore.
- **The intervention logic** – an understanding of what an intervention is expected to achieve and the assumptions behind how it is expected to do this. This is particularly important for: outcome and impact evaluations that make use of theory-based designs to understand causation; mixed-methods designs; and outcome-based approaches that look at contribution to results in multi-actor intervention or networked interventions. This may be revisited in the inception phase.
- **The availability of monitoring or other relevant data** or the possibility of generating such data within the resources allocated to the evaluation; also, the additional data required for the evaluation to answer the specific questions that commissioners and stakeholders have.
- **The conduciveness of the context** to carrying out an evaluation that looks at protection. This should include considerations around organisational 'climate' and leadership support to the evaluation; access, logistics and safety of the evaluation team; and ethical appropriateness. It also involves mapping the intended uses by the intended users.

11 The elements proposed could be used in general EHA work.

Figure 4: Evaluability analysis flowchart



Source: Adapted and expanded from Davies (2013: 24)

(*) **Note:** The 4 evaluability checklist are featured in the "TOOLKIT item #1 – Evaluability checklists for evaluation of protection in humanitarian action" on page 77 that accompanies this guide.

Ways to establish evaluability in evaluation of protection in humanitarian action

There are different ways of establishing the evaluability of a programme or intervention. In this guide we suggest two approaches: (1) Rapid evaluability scans, usually developed further during the inception phase of the evaluation; and (2) stand-alone evaluability studies. Many agencies may, however, undertake activities that lie between these two ends of the spectrum.

A **rapid evaluability** scan is a more informal and less structured analysis of evaluability that can flag issues that need to be expanded upon during the inception phase of an evaluation.

The objective is to clarify the protection-specific evaluability conditions of the intervention and help orient the design and methods for the evaluation. A rapid evaluability scan is done using modified and shortened versions of the evaluability checklists used for fully fledged evaluability studies.

Evaluability studies (sometimes referred to as Evaluability Assessments) are stand-alone exercises commissioned and carried out independently from an evaluation, before the decision whether to initiate an evaluation is confirmed[12].

Their objective is to help identify whether an intervention can be evaluated, and whether an evaluation is justified, feasible and likely to provide useful information. Their purpose is also to prepare the necessary conditions for the evaluation (UNEG, 2011: 17).

Carrying out a stand-alone evaluability study is recommended for evaluations emphasising protection-specific outcomes and impacts.

Commissioning fully-fledged evaluability studies covering protection can also be recommended for large-scale, inter-agency and high-stakes evaluations that have a primary line of inquiry (or area of focus) specifically on protection.

Table 1 summarises the trade-offs between different ways to establish the evaluability by clarifying the differences in timing, resource requirements and deliverables in rapid evaluability scans and evaluability studies.

12 For an example of fully-fledged evaluability study, see Davies et al. (2012).

Table 1: Overview on two approaches to establish the evaluability status

	EVALUABILITY STUDIES		RAPID EVALUABILITY SCAN
When	Commissioned independently from the evaluation as stand-alone exercises before the decision of initiating an evaluation is taken		Can be integrated into processes culminating in the inception phase once an evaluation is initiated
Approach	Desk phase + Field visit and stakeholder meeting	Desk-based only	Generally desk-based only Explore the same set of questions used for evaluability studies
Output delivered	Written evaluability report with recommendations to proceed, postpone or review evaluation plans	Written evaluability report with recommendations to proceed, postpone or review evaluation plans	Stand-alone evaluability report may not be required Evaluability considerations discussed with evaluation commissioning agency more informally Evaluability analysis feeds into the inception report with recommendations for the evaluation focus, sharpening of the evaluation questions, clarifying the variables and unit of analysis to improve the data collection plan etc.
Evaluability results' format	Evaluability checklist questions answered in detail + Scoring / adjusted scoring is produced to then aggregate in an index of the overall expected complexity of the evaluation	Evaluability checklist questions answered in detail + Scoring / adjusted scoring is produced to then aggregated in an index of the overall expected complexity of the evaluation	Evaluability checklist answered (checklist may be modified and shortened, depending on the evaluation and timing when the evaluability analysis is carried out)
Approx. length of the exercise	2 weeks	5-7 days	2-3 days should be factored in in the evaluation overall workplan/timeline

Costs and resources

Costs and resources can influence evaluation in all fields and disciplines. The evaluation of protection in humanitarian action is no exception. Resource requirements and cost considerations play an important role in the choice of the method and related tools used in an evaluation. However, there are generally too many variables involved in designing evaluations to provide actual cost comparisons for different evaluation approaches here (Rogers, 2011: 27).

The following factors are likely to determine the cost of an evaluation:

- Purpose and scope of the intervention
- Depth, thoroughness and date of usable context analysis and of the protection analysis (if one has been carried out)
- Number and type of evaluation questions being asked
- Expected generalisability of the evaluation findings
- Expected methodological rigour in the data collection and data analysis; also likely to be influenced by the qualitative or quantitative-leaning preference and orientations of the evaluation commissioning agency (Bamberger, Rugh and Mabry, 2012: Chapter 11)
- Choices related to carrying out primary data collection and ethical procedures (see Module B)
- Utility and reliability of existing monitoring data from the intervention being evaluated
- Complexity of the programme context (including access and security constraint expected to affected the evaluation fieldwork)
- Complexity and number of the interventions/components and sub-components that the evaluation is expected to assess and synthesise
- Ambition level regarding broad stakeholder verification and engagement in using the evaluation findings and conclusions
- Expected role of external evaluators.

SECTION 6: Evaluation questions and the intervention logic

Evaluation questions

Evaluation questions frame the focus of the evaluation and can help to tell a comprehensive story when the findings are presented (Kuster, et al., 2011: 40). In evaluations of aid interventions the most common framework for structuring evaluation questions are the Organisation for Economic Co-operation and Development (OECD) Development Assistance Committee (DAC) criteria, which focus on relevance, effectiveness, impact, efficiency and sustainability (OECD-DAC, 1991). These have been adapted to focus on issues arising in humanitarian action to consist of coverage/sufficiency, effectiveness, relevance/appropriateness, efficiency, connectedness, coherence and impact (based on Beck, 2006).

This sub-section applies these adapted criteria and considers how they can be adapted to the issues arising in evaluating protection in humanitarian action[13]. Definitions of protection used by different agencies will emphasise different criteria. The comments below are intended to contribute to reflection, recognising that their appropriateness may vary depending on the agency and above all else, on the focus of the evaluation itself. The choice of questions should ultimately reflect the purposes of the evaluation and its intended uses. There are a range of other frameworks for structuring evaluation questions that may be more or less appropriate for a given evaluation. For more information, see the ALNAP Guide to Evaluation of Humanitarian Action (ALNAP, 2016).

Coverage/sufficiency: Is the volume and distribution of resources sufficient to meet needs? To what extent are needs covered?

Humanitarian reviews often point out that in complex emergencies protection is the greatest need but that it receives far less resources than more visible, measurable and straightforward assistance responses (SOHS, 2015). Evaluating protection in humanitarian action is an important part of mapping coverage and sufficiency. Questions might focus on the specific operational environment of a given agency, or the 'big picture' of the extent to which protection efforts meet overall needs. The latter may include some critique of biases within the humanitarian system wherein protection needs that are difficult to measure and address, are given proportionally less attention than more straightforward relief assistance.

Effectiveness: How well were humanitarian objectives met? Was the response timely?

The most common and often dominant focus of aid evaluations in general is effectiveness, often framed by the term 'results', which may also encompass 'impact' (see below). Given the frequently large gap between needs and operational capacities, and the pressures on agencies to promise grand results, assessments of effectiveness may also include a measure of 'reality check' on the extent to which planned objectives were plausible. Further, the volatile context of humanitarian action in

13 It is notable that the OECD/DAC criterion of sustainability is missing in this list. It is here interpreted that factors related to sustainability in evaluating protection in humanitarian action are subsumed and somewhat nuanced in the criteria of connectedness and coherence.

general and protection in particular may generate a mismatch between rigid results frameworks and operational realities. Part of the 'reality check' is about providing input for learning how to adapt results to changing needs and operational opportunities. There may therefore be significant links to questions related to connectedness (see below).

Relevance/appropriateness: Do interventions address the priority needs of recipients? To what extent do they drive programme design?

Ideally, evaluators should be tasked with evaluating relevance against a pre-programme needs assessment and any further assessments undertaken in the duration of the initiative. A challenge facing evaluators may be whether and how to make up for insufficiencies in these assessments. Resources generally allow (at best) for a review of secondary sources of information about needs and all too seldom allow an evaluation team to gather empirical data directly. Such data may be essential in the case of an impact evaluation, with obvious implications for whether such an evaluation is viable.

In addition to the relatively little attention paid to protection noted above, perhaps the other great gap in humanitarian performance relates to engaging recipients in defining their own needs and programme design. The importance and contributions that this can provide are clear, as are the ethical imperatives. Methods are many. See "TOOLKIT ITEM #3 A partial menu of evaluation approaches and designs" on page 83. There are also ethical challenges and dangers, discussed in "SECTION 10: Ensuring that the evaluation is carried out in a protective and conflict-sensitive manner" on page 43.

Efficiency: Do outputs reflect the most rational and economic use of inputs?

While a seemingly straightforward criterion for evaluation, efficiency is also one of the most difficult and least developed in relation to protection. One reason for this is widespread confusion and misuse regarding terminology. The concept of "cost-benefit analysis" is frequently raised, but it comes from the field of economics and measuring. Measuring financial costs in relation to the value of the benefits of protection leads easily into comparing two completely different metrics (money versus human suffering or human dignity). Instead, rational and economic use may be best measured against a selected benchmark of similar programmes or activities and services.

A major focus in relation to efficiency is currently 'value for money', but even here it is essential to be clear about what 'values' are to be measured and how. Assigning such values to human rights violations is inevitably a contested notion.

Connectedness: Do humanitarian interventions take account of other key actors and efforts?

The humanitarian sector is coming under increasing pressure to strengthen links to other key actors and efforts (humanitarian and otherwise). Questions in this regard may refer to two broad categories. The first is coordination within the system. Whereas coordination of relief assistance has received a great deal of attention and has seen significant improvements in recent years, protection remains more contested. There have been major efforts, but it is too early to judge the overall results given ongoing disputes about mandates, relative roles and even basic definitions. For this reason, when initiating an evaluation it is extremely important to understand the way that a given agency frames the concept of protection and extent to which it strives to 'connect' with other actors.

Furthermore, considerable contextual analysis is required to understand which connections are appropriate, feasible and desirable. Ideally, this analysis will already have been undertaken by the agency, but this may not always be the case. (Re)constructing the contextual analysis underlying actions related to connectedness may be an important part of describing and critiquing an intervention's logic. This may include asking what the assumptions were about who would do what, and whether those assumptions were valid.

Central to this, protection issues are often very closely associated to power and the use of power in society – at the level of the state, the community and the household. Interventions in this area therefore inevitably become part of complex social processes which involve a number of actors. To be effective, the agency will have to make a contribution that is "connected" to a complex social context, and an understanding of this context is often a precondition for making any judgement about the value of the programme.

The other aspect of connectedness concerns to the extent to which an initiative has either aimed to enhance or even to relate to the broader protection environment. Any intervention should be designed to take into account the role of the state in protecting its population and respecting the rights of displaced populations. The extent to which an agency has the mandate, opportunity or ambition to enhance the role of the state or other national partners will vary. Here again, it is essential to clarify this when initiating the evaluation.

Coherence: Does the intervention adhere to core humanitarian principles and align with broader peace and development goals?

In many evaluations the analysis of connectedness will overlap with that of coherence. Reflections on 'who does what' will inevitably need to be anchored in an understanding of 'why they do it'. It may be assumed that there is broad consensus on core humanitarian principles and peace and development goals, but the interpretations may vary, as do the mandates and areas of engagement of different agencies.

Impact:

The definition of impact in humanitarian response is contested and often muddled with other criteria (SOHS, 2015). Given the considerable interest around impact evaluation, this is discussed in detail in "SECTION 8: Special considerations in deciding whether to undertake impact evaluations" on page 34.

Intervention logic and theory of change

Ultimately the evaluation questions should be selected to critically interrogate and unpack the initial intervention logic (often a 'theory of change') of the programme. In the inception phase evaluation teams will often engage with stakeholders to make an initial assessment of the 'formal' intervention logic, usually documented in a results framework (such as a logical framework) in order to clarify the assumptions behind the interventions in relation to all the above criteria. It is essential that this is not taken for granted, as many programmes, perhaps due to being designed in haste in response to a humanitarian crisis, lack a clear theory of change. Sometimes programmes have deviated from results frameworks that have been hastily submitted in order to obtain donor support without sufficient preparatory analysis. See "TOOLKIT item #4 – ODI/RAPID approach to Theories of Change" on page 94

Unpacking the logic or theory of change that has underpinned the intervention involves three fundamental steps:

- Making explicit the nature and scope of protection in humanitarian action on which the intervention was based
- Reconstructing how an intervention was expected to achieve its protection goals, taking into account the policy, institutional and conflict context
- Identifying the specific results the intervention intended to achieve from a protection perspective

Being anchored in the intervention logic means asking critical questions about the 'what' and 'why' of an intervention (project, programme, service, policy). We return to the issues surrounding intervention logics in "MODULE C – ANALYSIS" on page 62.

SECTION 7: Indicators

One of the most challenging aspects of planning an evaluation is to select rubrics See "TOOLKIT item #2 – Evaluative rubrics" on page 81 and indicators that are relevant for the evaluation questions. Ideally, these may have already been implicitly or explicitly identified in the intervention's results framework (at least regarding effectiveness) and relevant data monitored. In practice, however, initial results frameworks may not correspond to the issues of interest in the evaluation, or they may have been rudimentary and/or ignored in programme implementation. Weak 'ownership' of results frameworks is the first finding of many an evaluation, as the evaluators seek to determine what indicators the intervention has actually monitored and measured during implementation.

Indicators may also be selected as part of an evaluability analysis, but very often they are left to the evaluation team to select during the inception phase.

Evaluation questions require establishing indicators that measure quality and value. This is also likely to include an analysis of quality and value of the intervention as perceived by the participants, programme recipients and other key stakeholders.

The selection of indicators is often a de facto litmus test of the breadth of the consensus on:

1. What aspects of protection the intervention is expected to achieve.
2. How relevant the ambitions are in a given context.
3. The extent to which changes in the protection environment are seen as realistic and measurable within the timeframe of the intervention.

Relevant indicators of efficiency are notoriously difficult to determine in humanitarian interventions in general, but particularly in protection. As noted above, it is important to frame efficiency questions in relation to an appropriate benchmark.

At the stage of initiating an evaluation it is likely that some 'danger signs' will already be apparent regarding the data required for the selected indicators. The underlying assumptions on the feasibility of accessing the data and the expected level of confidence in it should be described in either the evaluability assessment or, more often, in the inception report. If necessary, alternative indicators may need to be selected. Some might be proxy indicators, as discussed in the following section.

Proxy indicators

When direct measurement is not possible, proxy indicators are used to approximate or understand the broader conditions that determine the likelihood that a change occurred. They are distinct from indicators that directly measure change. According to United Nations Development Programme's definition, proxy indicators do not indicate that change has occurred but rather suggest that there

conditions are present which are conducive to the desired change (Corlazzoli and White, 2013: 20). As such they may provide evidence of the relevance of the programme in addressing protection needs, even if they cannot verifiably judge the actual effectiveness of the intervention.

In EHA in general, and in evaluating protection in particular, proxy indicators are helpful in situations where regular data collection mechanisms are insufficient, when monitoring mechanisms are disrupted, and in situations where certain lines of inquiry for primary data collection are not feasible or appropriate (or even ethically defensible, as discussed in Module B).

Proxy indicators also offer a way to measure more abstract concepts, such as well-being, trust or community cohesion. When looking at these aspects of performance, such indicators will be important in bringing contextual factors into focus, but there are certain pitfalls.

» When proxy indicators make use of secondary data – such as administrative data and data logs from helpline, injury surveillance and health centre data, for instance – they are prone to **bias from usage**: they only capture cases that have been reported or detected or for which services were sought.

EXAMPLE: Use of administrative data as proxy indicators

Data from health centres is potentially useful for triangulation or as a proxy indicator, especially if they cover a particular response linked to child protection. For example, the 2009 Kenya situation analysis used data from the Gender Violence Recovery Centre of the Nairobi Women's Hospital, and triangulated it with survey data, caseload reports and official reports of the Kenya Police to establish changes in violence over time.

Data from child helplines can be useful, for example, to triangulate information from other sources like surveys and police data. In the absence of any other data, they could also be reviewed for a basic, highly aggregated analysis and/or to establish trends over time.

However, it is important to note that this type of data can be biased from usage because it captures only those cases that have been reported or detected or for which services were sought.

Source: UNICEF (2015: 46)

EXAMPLE: Indicators used for hard-to-measure issues

In a collaborative approach to psychosocial programming, a number of Palestinian agencies agreed to specific indicators of aspects of psychosocial well-being: for example, reduction in troubling dreams (as a measure of emotional well-being) and increasing collaborative behaviour with teachers and peers (as a measure of social well-being).

Identifying such indicators has enabled psychosocial workers to gather clear results on their interventions. It has also led to more awareness of these aspects of behaviour among teachers, and to an increase in parental involvement in children's activities at school.

Source: IFRC (2009: 175) cited in Ager et al. (2011)

Proxy indicators are particularly important where the evaluation involves measuring an impact in terms of things that did not occur, for example human rights violations or incidents of gender-based violence (GBV). For example, measuring reductions in GBV in general may provide a proxy measure for assessing efforts to strengthen the protection environment and the role of the state. An evaluation can measure relevant trends as a proxy indicator, but it should be recalled that these trends usually cannot be directly attributed to the intervention.

In the conflict transformation and security and justice sectors, one area of emerging evaluation good practice is around the use of bundles (or baskets) of proxy indicators to help measure broader trends by looking at nuances of change (Corlazzoli and White, 2013). For example:

- In Bangladesh, hospital admission records have been used to verify media reports of acid attacks against women.
- In Afghanistan, parental perceptions of safety (garnered through focus group discussions) were combined with school attendance data to determine trends in freedom of movement.

As highlighted in the Afghanistan example, data from another sector can be useful. School attendance is becoming widely used as an indicator of freedom of movement. In the same vein, indicators relating to patterns of participation in the public sphere (such as cultural, religious events, or weekly markets) are also increasingly used as part of the 'basket of indicators' around freedom of movement.

Box 2: Criminal justice data and public health sources

There are two very different sources of official data used by conflict crime and violence programming that are relevant to protection programming monitoring and evaluation – criminal justice based data sets, and public health based data sets:

Criminal justice sector – collects primary data categorised as crime in the respective legal system. Recording can take place in various places – e.g. recording a homicide could occur (a) as a body is found (police data), (b) as it is autopsied (forensic data) or (c) as criminal prosecution ends in a judgement (ministry of justice data).

Public health sector – collects primary data on violent deaths as they occur in hospitals or health care (e.g. intensive care units), or as deaths are recorded in national vital registration statistics (usually under the ministry of health).

The key difference is the focus on **events** (e.g. police records of the number of crimes) or on the victims (e.g. emergency room services), although the distinction is not absolute. The availability and quality of the data varies widely, and some argue that homicide data is the strongest, because as a very serious offence it is more likely to be recorded in criminal justice records. Conversely, public health data may have a better chance of picking up on proportions of and effects of violence (but not in the case of mental health, as provision tends to be poor).

Some secondary data sources, such as observatories (or Armed Violence Monitoring Systems) combine both criminal justice and public health sector data. The Jamaica Crime Observatory, for instance, maps data from the Jamaica Injury Surveillance System onto police crime data. State capacity to compile data sets in both criminal justice and public health may however be weak.

Source: Small Arms Survey (2013)

EVALUATOR'S INSIGHT: Thinking outside the box to gather data

Creativity and imagination can generate new data collection techniques for evaluative analysis.

For example, an evaluation that sought to determine the protective benefits of World Food Programme food relief in Darfur used proportional piling of beans to understand sources of livelihoods, and whether people had to take risks by leaving IDP camps to seek work or gather firewood. Interactive theatre can also be used as an evaluative measure, with drama stopping at critical junctures in the play and the audience actively choosing how a story should proceed, and recording those decisions. However this audience input is a public statement, and so caution is needed to ensure that people taking a public stand are not put at risk.

SECTION 8: Special considerations in deciding whether to undertake impact evaluations

Impact evaluation is not a universal answer to the challenges evaluating protection in humanitarian action and certain issues should be explored before embarking on this approach. Four starting points could help thinking about this specific type of evaluation:

1. Clarify the scope of the term 'impact'
2. Consider what learning and accountability needs can justify what may be a costly impact evaluation
3. Clarify the level of certainty that the evaluation expected to achieve
4. Consider whether the resources and contextual conditions are sufficient to undertake an impact evaluation with an acceptable level of rigour.

The term 'impact' is often understood and used differently by evaluation and programme actors, with different nuances and meanings attached to it[14]. The State of the Humanitarian System report (2015) notes that humanitarian evaluations often conflate the term with various aspects of effectiveness and sustainability. Stern has flagged (2015: 8) that 'impact' in an evaluation can be seen as covering:

- The effect as intended by policy-makers and programme planners or as experienced by intended aid recipients and others
- An immediate experience or a more enduring change in circumstances or capacities
- The effects at the level of individuals or communities or institutions.

In a strict sense, the term often associated with extent to which the initiative has impacted on the conditions of the affected population. It may also include unintended negative impacts on their well-

14 The ALNAP EHA guide defines impact as the wider effects of the programme – social, economic, technical, and environmental – on individuals, gender- and age-groups, communities and institutions. Impacts can be intended and unintended, positive and negative, macro (sector) and micro (household, individual), short or long-term. .

being or the range of impacts on different sectors of the population. On the other hand, impact is also often used to refer to the outcomes of an intervention in relation to changes in the way services are provided, the readiness of the state to shoulder its protection responsibilities, or other broad 'results'. Most agencies (and their donors) have explicit or implicit understandings of the scope of the term. These interpretations need to be clarified when assessing evaluability.

Once definitions have been clarified, it is essential to look at the extent to which impacts (in whatever sense) can be attributed to the intervention. Depending on the expected sphere of influence of the intervention, attention may focus more on outcomes in terms of overall access to services, government commitments or other institutional changes. The latter may be related to changes in the capacities of, for example, national partners in government and civil society to undertake protection responsibilities. It is therefore important to anchor this analysis in a thorough mapping of stakeholders and their diverse interests and capacities to influence different processes and actors. See "TOOLKIT item #5 – ODI/RAPID influence and interest matrix" on page 95 Issues related to the spheres of influence and interest are discussed further in "SECTION 11: Data sources and constraints in data gathering for evaluating protection in humanitarian action" on page 50 below in relation to attribution.

Considering the challenges inherent in designing and carrying out impact evaluations – particularly in humanitarian contexts[15] with their specific resource and expertise requirements – one question to ask is when there are sufficient grounds to justify this specific type of evaluation. We suggest (drawing on work by Rogers, 2012 and Chigas, Church and Corlazzoli, 2014) that at least one of the following conditions should be met in order to justify the investment (time, technical, financial) needed to initiate, resource, design and carry out a credible, high-quality impact evaluation:

a. The intervention is significant enough (in terms of size, policy prominence or potential consequences) to justify such an evaluation.
b. The intervention is strategically relevant vis-à-vis the context, conflict and/or protection analysis against which the intervention (ideally) is situated.
c. There is limited, untested or contested evidence of 'what works, for whom, and where' that the impact evaluation would look at. For example, there would be differing views about whether a specific type of intervention is effective in the context and for a given target group. In such cases a somewhat more modest approach looking at plausible relevance, connectedness and coherence of a given intervention model from the perspective of different stakeholders may be sufficient.
d. There is access to the right range of actions by different actors. That is, the evaluation will need to have the capacity and mandate to investigate the actions of a range of agencies whose actions are required to achieve the intended impacts.
e. The intervention is ready for impact evaluation in terms of there being sufficient baseline data and sufficient time passed since the start of the intervention to provide a basis for measuring change.

15 See for instance Cosgrave and Buchanan-Smith, 2013: section 1.5.

Finally, as discussed in "MODULE B – Data collection in evaluating protection in humanitarian action: Practical and ethical implications" on page 42 collecting primary data about protection-related impacts with respect to incidents and incidence prevalence requires confronting several profound risks and ethical issues. Evaluating the impact of protection initiatives may require the collection of such data. The decision to undertake such a course of action requires careful judgement and possibly considerable risk mitigation efforts.

SECTION 9: Selecting approaches, designs and methods

Every evaluation requires an overall approach, and within that a design and methodology that is: (1) in line with its purpose and questions; and (2) responsive to the features of the programme, the specific programme components and sub-components being examined, and data and contextual factors.

This section of the guide offers an admittedly limited overview of what is likely to be a broad menu of options for selecting overall approaches and designing methods for evaluating protection in humanitarian action. Ideally, the richer the menu[16], the broader the options in the evaluators' armoury to come up with a design and specific tools that can make protection work more responsive to diverse and complex programme features and to prevailing data and environmental constraints. These include the:

- Nested nature of protection actions and different strands of work that co-exist in protection in humanitarian action
- Likely presence of spill-overs between different lines and modalities of protection work
- Heightened requirements for ethical and conflict sensitivity considerations in the evaluation process, particularly in the data collection stage

The basic terms relating to evaluation approaches, designs and methods used in this section are clarified in Box 3.

16 Increasingly, design and methodological pluralism in evaluation are put forward in general aid evaluation theory and practice as critical factors to drive improvements in the way evaluations can become more responsive to challenging external environments where the programmes are being carried out, and to complex programme features that are required in this type of environments (e.g. decentralised programmes; collaborative and network-based actions; nested interventions; programme with distributed governance).

Box 3: Key concepts and terms relating to evaluation approaches, design and methods

Key terms relating to evaluation design and methods are often used differently by different authors in different evaluation manuals. In this guide we suggest using the following:

EVALUATION APPROACHES are usually the largest category in an evaluation typology. They are defined as the overarching set of principles and orientations guiding the design and subsequent implementation of the evaluation. Evaluation approaches are then pursued through a range of data collection and analysis methods. Examples of approaches include contribution analysis (see "SECTION 14: Analysing causality, attribution and contribution" on page 68), or theory-driven evaluation (built around analysing the theory of change, see "TOOLKIT item #4 – ODI/RAPID approach to Theories of Change" on page 94).

EVALUATION DESIGNS are the methodological blueprint of an evaluation. They embody the logic that guides how an evaluation is conducted and guide the reasoning required in an evaluation to draw specific conclusions to answer the initial questions asked. At the broadest level, a design should consist of four elements: questions, description of the theory or hypotheses that underpin the evaluation, selection of data sources and the use of data (King, Keohane and Verba, 1994). The major elements of an evaluation design are:

- The unit of analysis that will be employed by the study and how they will be selected
- The parameters, or aspects of the programme/intervention that will be evaluated
- The comparisons needed – if any
- The variables, indicators and concepts being measured
- The boundaries of the evaluation (e.g. time, population, geography) – decisions around boundaries will affect the generalisability (external validity) of the findings
- The level of precision needed to produce useful and credible evaluation results (Rog, 2005: 114-115).

In terms of quality, sound evaluation designs should anticipate limitations and challenges, and be systematic in terms of being transparent with regard to original intentions (e.g. at evaluation inception stage) and actual implementation (Yarbrough et al., 2011: 201).

EVALUATION METHODS are used in this guide to refer to the tools and techniques which can be utilised in support of an evaluation approach, while **EVALUATION METHODOLOGY** indicates a procedure or system by which evaluation methods are organised (White and Phillips, 2012: 4).

The purpose of the evaluation and the primary questions it asks provide the point of departure for the subsequent choice of evaluation design. Other factors that help shape the design include[17]:

a. The extent to which the programme was designed around a given monitoring approach anchored in a certain type of indicators (e.g. outcome mapping)
b. Expected uses of the evaluation results
c. The resources available for the evaluation
d. Overall feasibility concerns, including access, logistics, security for the evaluation team and for those included in the data collection
e. The level of certainty the evaluation expected to achieve (i.e. the level of certainty the evaluation commissioning agency and the intended users of the evaluation need to have in order to be confident in using the findings)
f. Whether the approach and design of the evaluation can be implemented with integrity given ethical considerations or constraints on access or resources
g. The methodological preferences of clients, commissioning agencies, other stakeholders and the evaluators themselves, often related to what types of data they trust and what approaches 'push the right buttons' in their respective organisations
h. The availability of monitoring and secondary data – this is critical in evaluating protection in humanitarian action considering the more restrictive options for primary data collection and for group-based data collection on sensitive issues
i. The time allocated to the evaluation.

The level of precision, generalisability and utility of the findings of any evaluation is affected by the representativeness of the sample and how the data will be analysed:

- How units are sampled for the data collection step in the evaluation
- Whether or not the design chosen makes use of some form of comparison (Rog, 2015: 115).

Decision-makers are often prepared to make decisions if they are, for example, 70-80% certain of the accuracy of the evidence provided. Different contexts and different types of decision call for different thresholds of certainty. 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 stakeholders (Davidson, 2005: 69). This, in turn, also relates to the extent to which the evidence can show clear contribution or attribution given the spheres of control, influence and interest of the intervention (see "TOOLKIT ITEM #3 – A partial menu of evaluation approaches and designs" on page 83 and "TOOLKIT item #5 – ODI/RAPID influence and interest matrix" on page 95)

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 will facilitate any up-front discussions about the trade-offs between budgets, time lines, and the breadth and certainty of conclusions (Davidson, 2000: 25).

17 The factors suggested here draw from Rog (2005: 114-116), Davidson (2005) and Bamberger, Rugh, Mabry (2012: 225 and 330-334).

Having some clarity on stakeholders' expectations should thus help evaluators better explain – and sometimes defend – their design and method choices.

Evaluative thinking is moving away from preconceptions about a single 'gold standard' and a presumed need for quantitative impact evaluations to supplant the gold standard with a new platinum standard: methodological pluralism and appropriateness. (Patton, 2014 b) In this light, methodological pluralism and appropriateness should inform decisions on the most appropriate blend or best fit of designs and methods to answer the evaluation questions at hand (Alexander and Bonino, 2015: 13-14). Selecting evaluation approaches require clarity of purpose, attention to process, and coherence within the overall evaluation plan[18].

Selecting an evaluation approach provides a general orientation, but does not automatically predetermine data collection or data analysis methods. For example, key informant interviews could be part of any approach but are not required by any single one. Similarly, gap analysis could be used to analyse data regardless of the approach and is not a prerequisite for any particular evaluation approach (Rogers, 2012).

Given that no evaluation approach can cover all needs, most combine a variety of methods. Some of the main reasons for combining and mixing methods include the opportunity to:

- Compensate for the weaknesses inherent in a given evaluation design and method
- Increase the credibility of evaluation findings when information from different data sources converges (triangulation)
- Deepen the understanding of the programme/policy, its effects and context, including the potential for generalising the findings and conclusions (Bamberger, Rugh, Mabry, 2012).

There are three main types of mixed method design:

» Sequential mixed method: Quantitative method followed by a qualitative method or vice versa

» Parallel mixed method: Quantitative and qualitative components conducted at the same time

» Multi-level mixed method: Where a large evaluation is conducted at multiple levels, with both quantitative and qualitative approaches being used at each level, including analyses of both direct protection actions and the protection environment.

Further, different methods will usually be required depending on whether the design will be single level (e.g., the household, organisations or institutions) or multi-level (e.g., a country programme that requires description and analysis of links between different levels) (Peersman, 2014: 4). Regardless of the specific mix or preference for a given set of data collection and analysis tools in an evaluation, there are a number of techniques – some of which are in the table below – that can strengthen the evaluation design by bridging the use of qualitative and quantitative data collection and analysis methods (Table 2).

18 This section of the guide draws from Stufflebeam and Coryn (2015); Patton (2015); Church and Rogers, 2005; Stern et al. 2012; Bamberger, Rugh, Mabry, 2012; Chigas, Church, Corlazzoli, 2014, Rogers, 2012; Tsui, Hearn and Young, 2014.

Table 2: Techniques that can strengthen mixed-method design

Technique	How it bridges qualitative and quantitative methods
Triangulation	The combination of data, analysis and findings generated from qualitative and quantitative analysis can increase the strength of the causal inference. In general, the validity of the evidence generated using triangulation is enhanced when two or more independent estimates can be compared. Both quantitative and qualitative methods use triangulation. But often in slightly different ways for example: • Quantitative methods are usually more concerned with using triangulation to check the internal consistency of measurements (e.g. for survey instruments) • Qualitative methods prioritise the use of triangulation to verify the information gathered and deepen and broaden understanding of an issue or phenomenon through obtaining multiple perspectives from different sources. This is done for example when the use of purposive sampling requires rich explanations and narrative that can illuminate cases of interest.
Process tracing / Process analysis	Qualitative analysis focused on processes of change within cases may uncover the causal mechanisms that underlie quantitative findings. Without process analysis it may not be possible to assess whether failure to achieve a certain outcome is due to design failure or to implementation failure. Most useful as a method for identifying, testing and validating hypothesised causal mechanisms within case studies. Examination of multiple cases may be used to build up a body of evidence.
Focus on tipping points	Qualitative analysis can explain turning points and crucial junctures for change within quantitative time series and changes over time in causal patterns established with quantitative data.
Using quantitative data as point of departure for qualitative research	A quantitative data set can be used as a starting point for framing a study that is primarily qualitative.

Sources: Tarrow (2009, reproduced in Stern et al., 2012: 33), Bamberger, Rugh, Mabry (2012: 229; 326-330), White and Phillips (2012).

EXAMPLE: Mixed evaluation method for UNICEF's child protection response to the 2004 tsunami in Indonesia

The evaluation, which was commissioned by UNICEF's Child Protection Department, aimed to determine the impact of UNICEF's response to the South Asian tsunami within the child protection sector. It followed the evolution of the three child protection work strands of the programme in Aceh (children without family care, children without psychosocial support, and victims of exploitation and abuse).

The evaluation employed a sequential mixed methods approach to combine comprehensive coverage with in-depth analysis. It focused on three districts to compare results between tsunami-affected and mainly conflict-affected districts, which allowed for comparisons between those areas with a strong operational UNICEF presence and those areas with less. The evaluation design also compared different interventions with one another or, where a similar programme did not exist, with groups of children who did not receive the intervention. For more, see www.unicef.org/evaldatabase/index_59604.html.

Source: UNICEF (2015: 201)

MODULE B – Data collection in evaluating protection in humanitarian action: Practical and ethical implications

» This module is primarily addressed to evaluation teams.

Content of this module at a glance		
	What is this section on?	Who is this section for?
SECTION 10:	Provides guidance for how to ensure that evaluations are carried out in a protective and conflict-sensitive manner	Primarily for evaluation teams
SECTION 11:	Delves specifically into how to select data sources and address constraints in data gathering	Primarily for evaluation teams
SECTION 12:	Explores how to approach data gathering on less tangible dimensions	Primarily for evaluation teams

The choice of the most appropriate data collection approaches and tools for evaluations in humanitarian contexts has practical, protective and ethical implications. The stakes are likely to be high due to the challenges and constraints described above, as well as to the heightened need for ethical safeguards and protocols to inform data collection.

This section starts with an overview of the challenges and constraints in collecting data for evaluating protection in humanitarian action. It then presents some ethical and practical implications that should be considered when taking decisions about:

- Which type of data can be collected for which purpose
- From which sources and from whom the data should be collected
- Minimising possible negative and harmful repercussions relating to data collection.

SECTION 10: Ensuring that the evaluation is carried out in a protective and conflict-sensitive manner

When evaluators are asked to look at humanitarian protection, it is critical to infuse the evaluation process itself with a 'protection perspective'. This is relevant to all evaluations carried out in conflict and insecure settings, or other situations of violence. The starting point – as with general EHA practices – should be a consideration of 'do no harm' vis-à-vis those taking part in, or being consulted during, the evaluation process, in particular during fieldwork and for data collection purposes[19]. (See Box 4 for more on the 'do no harm' concept).

Moreover, carrying out an evaluation in a conflict-sensitive manner should not be seen as only relevant for those evaluations explicitly examining whether an intervention has had any interaction with conflict[20]. Evaluators should be aware of how any evaluation could contribute to exacerbate tensions and compromise future access to affected populations. For example:

- Expectations may be raised that by taking part in the evaluation (e.g. during data collection) more aid will be provided, which could result in frustration
- The views of individuals and groups that are most at risk may be exposed, making them more vulnerable to reprisals by powerful actors
- The evaluation team could be perceived as gathering intelligence for one of the parties in conflict
- Focus group discussions (FGDs) could become heated, or conversely reinforce divisions by missing opportunities to bring groups together in FGDs
- The evaluation could become part of the battlefield for public opinion – people may respond in ways intended to promote a given agenda, raising concerns about both the credibility of findings and the ways that an evaluation report will be used in the future
- The evaluation could present a biased analysis if it does not adequately present different stakeholder views
- A predominance of views from more powerful/accessible informants may reinforce patterns of inequality and marginalisation
- Contested conclusions or recommendations may contribute to increased tensions.

Undertaking an evaluation in a conflict-sensitive manner involves:

- Assessing whether any steps in the evaluation process could contribute to tensions (this will need to focus on data gathering, analysis and dissemination of the report in particular)
- Carrying out new (or updating existing) conflict, context and stakeholder analyses, as this will inform the sampling frame and help identify possibly bias in the evaluation

19 These points are also covered in the United Nations Evaluation Group (UNEG) guidance on ethical obligations to those to initiate, manage and carry out evaluations (UNEG, 2008). These obligations include: respect for dignity and diversity; human rights; confidentiality; and avoidance of harm.

20 This is a core area of evaluative work in the realm of peace-building evaluation, evaluation of conflict transformation and evaluation in the security and justice sectors.

Box 4: What do we mean by 'do no harm' and 'Do No Harm'?

The term 'do no harm' can be confusing as it has different connotations in the humanitarian and conflict fields.

The principle of doing no harm derives from the medical principle that the physician should cause no harm to the patient. It has been adopted and adapted in other fields.

Definition 1 – From a humanitarian perspective 'do no harm' is a widely used term but is often not well defined. Within the Sphere Handbook (Sphere Project, 2011) it is captured under protection principle 1: 'avoid exposing people to further harm as a result of your actions', which includes not only violence and rights abuses, but also physical hazards. In common practice 'do no harm' has sometimes been used to mean avoiding or minimising any adverse effects from an intervention: for instance, siting of a latrine too close to a well.

Definition 2 – From a conflict sensitivity perspective Do No Harm (capitalised in this guide) refers to a specific 7-step framework that can be used to assess the conflict sensitivity of an intervention. It was developed by Collaborative for Development Action (now CDA), and is one of the most widely rolled out 'tools' for conflict sensitivity (see www.cdacollaborative.org). Conflict sensitivity means ensuring that an intervention does not inadvertently contribute to conflict, and where possible, contributes to peace (within the confines of an organisation's mandate). In this framing, Do No Harm relates only to conflict-related risks. This is commonly broadly defined to include many protection related risks, as there is a significant overlap between conflict and protection related risks.

The field of conflict sensitivity is of course broader than the Do No Harm Framework – there are many other tools in the conflict sensitivity toolbox and there is much practice and analysis that relates conflict sensitivity in a more comprehensive perspective, beyond tools and project modalities. For further information on the Do No Harm approach see http://cdacollaborative.org/wordpress/wp-content/uploads/2016/02/From-Principle-to-Practice-A-Users-Guide-to-Do-No-Harm.pdf

Safeguards and ethical considerations in data collection

Several of the points above overlap with ethical considerations in human subject research more generally[21]. These strongly apply when information on sensitive issues is sought directly from individuals who may have suffered harm or abuse, particularly in the area of sexual and gender-based violence (see Box 5).

In addition to 'do no harm' considerations, **data protection, confidentiality and informed consent** are essential safeguards that should be put in place before entering into any primary data collection

21 In the context of evaluation, ethics has been defined as encompassing concerns about the rights, responsibilities, and behaviours of evaluators and evaluation stakeholders (Yarbrough et al., 2011: 106).

activity specific to evaluating protection in humanitarian action[22]. To ensure confidentiality, data storage may require encryption software and due caution with regard to use of remote-access databases. Collecting information from individuals and stakeholder groups can create risks of reprisals for people disclosing information about their experiences. Focus group discussions should only be used where people have a common experience.

Confidentiality and its link to **data protection** are part of the ethical repertoire of the evaluator. They refer to the procedural devices to assure the privacy of the respondents during the data collection phase. The general principles underpinning data protection and confidentiality in evaluation is that people 'own' their own life experiences and that attributable data is only available to the evaluator on a negotiated basis (Kushner, 2005: 74).

Box 5: Special considerations for primary data collection on sexual violence

The World Health Organisation's Ethical and Safety Recommendations for researching, documenting and monitoring sexual violence in emergencies (WHO, 2007) makes the point that 'basic care and support for survivors must be available locally before commencing any activity that may involve individuals disclosing information about their experiences of sexual violence'. Evaluations are often assessing the availability of such services, and therefore these conditions can rarely be guaranteed in advance. Reflecting on the implications of this, Bain and Guimond (2014: 16) conclude that prevalence studies and other types of data collection conducted in the absence of gender-based violence services should be seen as in violation of humanitarian ethics.

Any quotes used in the evaluation report should be anonymised and completely unrecognisable regarding whom they came from, although some contextualisation can be given for the comments (e.g. 'An elderly widow from Damascus said…'). A report may be sanitised and circulation limited to avoid spread of information that could be associated with certain informants, even though with digital copies and files some risk of leakage remains.

Extra care is also needed in data recording and data storage, as there is a risk of hacking, confiscation or data theft. Time should be invested upfront to establish the recording and storage system. A good practice from the handling of research involving HIV is relevant here – to anonymise cases and keep only random numbers on the files, with names linked to numbers stored elsewhere. In a survey carried out in Nepal, all names were removed from all materials and kept in strictly confidential controlled files, while call records were kept in a separate place.

22 In many respects, the Child Protection community has paved the way in researching and clarifying informed consent procedures when working throughout the programme with children including in emergencies and crisis settings. For example, the UNICEF released an online resource portal called Ethical Research Involving Children (ERIC) (http://childethics.com) to compile and make accessible the latest resources and expert thinking about key ethical issues involving children and how these might be addressed in different research (and evaluation) contexts.

Extra security (encryption) measures may be needed for computers linked to the internet, particularly where data is stored and shared via remote log in.

Ensuring informed consent

Ensuring informed consent is about ensuring that people understand why they are being asked questions and how their answers will be used (see ICRC, 2013: 93). Programme participants and other stakeholders contacted as part of an evaluation should freely consent to participate in the exercise without being pressured to do so. Informed consent also includes reassurance that declining to participate will not affect, for example, any services provided to those who prefer not to participate (Brikci and Green, 2007: 5). Obtaining informed consent before proceeding with gathering information requires evaluation teams to:

- Ensure that all potential respondents, including children and young people, fully understand what is involved in their participation
- Encourage questions and clarification before proceeding with interviews or other data gathering exercises
- Allow sufficient time for potential participants to reflect on and decide about taking part and assess the respondents' understanding of consent by, for example, using quizzes or asking questions one-to-one or asking them to summarise what they have been told
- Equip interviewers with information on services available to the interviewees (e.g. health and social services)
- Be aware that some special considerations apply for all data gathering exercises expected to cover issues or experiences relating to sexual violence (See Box 5)
- Let the participants know that they can withdraw from the data collection exercise at any time.

There may be cases where informants are suspicious of the evaluation team or data collectors, or where they may only comfortable talking to outsiders rather than with people in greater proximity to their lives. Key informants might not agree to sign any consent documents, believing that signing a document is a trick. The act of asking them to sign may stop them from engaging at all.

Box 6: Informed consent process before interviews – the HESPER Scale example

The HESPER Scale is a tool developed by WHO with King's College London to look at perceived needs of people in crisis contexts across 26 'need items'. Ratings are made for each need item according to unmet need (or serious problem, as perceived by the respondent), no need (or no serious problem, as perceived by the respondent), or no answer (i.e. not known, not applicable, or answer declined). Respondents are also asked to name any other unmet needs not already listed.

The administration of the HESPER Scale by interviewers to respondents is preceded by an informed consent process. This is to ensure that respondents take part in the interview voluntarily, without coercion or fear that they will miss out on benefits if they do not participate, and to help avoid raising unrealistic expectations. Informed consent may be taken either verbally or in writing, depending on the context.

At a minimum, this should involve explaining to the respondent who the interviewer is and the agency he or she represents, the reasons for the survey, and an overview of the interview process, including the amount of time needed. Furthermore, it should be clarified that participation is anonymous, completely voluntary, that no compensation will be paid, and that there will be no benefits to respondents if they participate.

The interviewer should then answer any questions the respondent may have, before asking whether the respondent agrees to take part.

Ideally each respondent should be given a participant information sheet explaining all of the above (which they may either read themselves, or which may be read out to them), and each respondent should sign two copies of this sheet (one for the respondent to keep, one for the interviewer) as consent to take part in the survey. If the respondent does not agree to take part, he or she should not be pressured into doing so. Respondents should also have the right to withdraw from the interview at any point without having to give a reason.

Source: WHO and King's College London (2011: 24)

Pay attention to who collects the data

In evaluating protection in humanitarian action, it is critical to pay attention not only to how is data collected, but also by whom.

Recruiting local researchers (not just from the country but from the locality of the research) can enable access to data that would be too risky to collect. However, this also reflects a transfer of risk – where it is delegated to local partners and individuals. Ethnicity, gender, caste, religion and other factors can generate risks for both the interviewer and the interviewee.

All data collectors should receive substantial training in interviewing techniques, including knowing when to stop an interview (for instance if someone else has entered the room or informant is deeply distressed), being able to identify if approaching an informant would put the informant at risk, and being able to determine an appropriate place to conduct an interview.

While expertise in evaluation and protection is essential, those with long experience in the geographical area are particularly crucial to evaluations in conflict contexts. They have the contacts and are more likely to get access, and to hear the real voices of the people on the ground. They can identify and navigate the stakeholders, contextualise informants and their biases, and distinguish exaggeration from fact. They can provide essential insights into the political economy of the institutional environment in which the intervention is being implemented.

Dealing with sensitive data: legal implications and political sensitivities

Data collected can have legal, prosecutory and disciplinary ramifications: Sensitive protection-related data, gathered as part of an evaluation exercise, can have legal and disciplinary implications. For example, there could be cases in which the data gathered as part of an evaluation points towards criminal activities, violent acts liable to prosecution under national legislation, or malpractice and abuse (including sexual abuse) on the part of agency staff or partners. The obligations on the evaluator in responding to these situations will differ depending on the nature of the information, the jurisdiction in which alleged activities occurred, and the policies of the agency concerned. In all cases, it is important that data collection protocols clarify at the outset the evaluators' options and channels when uncovering this type of information.

Data can touch on issues which are politically sensitive: A main challenge here is that evaluators may not be able to use all the information gathered during the fieldwork in their report. When drafting the evaluation report, this may result in some gaps in the chain of argument and chain of evidence used to substantiate findings that were generated drawing from sensitive information. One report from Oxfam touches on this specific point:

> Due to its sensitive nature, some of the material collected for this review has not been included in this report. The conclusions and recommendations aim to elicit important learning from the full range of experience including that which is not documented here, and the reader may find some disjuncture in making direct links from case studies to some conclusions and recommendations. (Oxfam, 2011: 37)

Evaluators should ask whether the programme or agency has established some guidance on how to deal with sensitive data in terms of mandatory reporting and disclosure of sensitive data in the final evaluation report.

When ethics and transparency protocols are discussed for evaluating protection in humanitarian action, there should be space to clarify the extent to which evaluators are subject to mandatory reporting.

EXAMPLE: Clarifying mandatory reporting for evaluators when gathering sensitive data

Below is an excerpt from the ethics protocol used during recent fieldwork conducted in Rwanda by UNHCR, CPC and AVSI to develop a child protection index:

> 'AVSI and UNHCR Rwanda agreed to exempt researchers and data collectors from any existing mandatory reporting policies of abuse and violence. When a case was identified, the respondent was informed of services, and asked if s/he would like assistance in accessing those services.' (Meier, Muhorakeye and Stark, 2015: 35)

One of the recommendations following the completion of the first round of field-based data collection, the research team suggested making some revisions to the exclusion of mandatory reporting, particularly in the case of suicidal ideation (92).

SECTION 11: Data sources and constraints in data gathering for evaluating protection in humanitarian action

Data-gathering efforts should be informed by awareness of a range of risks and constraints, some of which can be planned for and mitigated. At the very least, potential scenarios should be considered when planning data collection and a 'do no harm' approach should be applied. The following factors should be considered in making decisions about sources and constraints:

Insecurity and constrained access: Insecurity means that evaluators tend to have short visits in easier-to-reach places. Sometimes interviews cannot be pre-arranged as this would create risks, so the evaluator can only speak to whomever happens to be there at that time (a 'convenience' sample). Informants often distrust outsiders and are reluctant to talk. This makes it harder to draw out data, and also to achieve and document 'informed consent'. Interviews at places where people gather (e.g. markets) in more secure areas may somewhat reduce these risks.

Access of international members of the evaluation team may be severely restricted: This can in turn lead to reliance on more junior evaluation team members with little experience or training.

Trauma and shame: Asking people to describe traumatic experiences can re-traumatise them. They may feel shame for the experience, particularly where they have experienced sexual violence. As noted in the sub-section on special considerations in primary data collection below, alternatives to collecting data from traumatised individuals should be considered.

Creating or aggravating risks for informants: Informants may hold a well-founded fear of reprisal for disclosing information about their experience. This relates not only to what is written in the report, but also to secure storage of data (leading to requirements for data encryption). The conclusions or recommendations of the evaluation report, if made public, could inadvertently contribute to tensions and thus increase vulnerability. Even if transparency is normally a principle to strive for in evaluation, the special circumstances surrounding protection may suggest that some reports must remain confidential.

Poor data environment: Even where there is a baseline and indicators, a rapidly changing environment (including rapidly fluctuating populations due to displacement and/or cross-border movements) may mean that baseline data and indicators are no longer relevant, or that indicators have not been regularly monitored. Official records including national statistical data or secondary sources are often weak to non-existent, and poor or non-existent national statistics affect the choice of sampling frame. The political sensitivity of key variables may have prevented data from being collected or negatively affected its credibility. There may also be challenges in determining what makes a 'typical' case for case study selection.

Polarisation: In situations of conflict, views tend to polarise, making the risk of bias is very high. The evaluation itself can become part of the battlefield for public opinion as informants or stakeholders respond and act strategically – trying to use the evaluation to support a particular policy narrative regarding the causes of conflict or donor responses (sometimes referred to as 'policy-based evidence formation'). Similarly, there can be issues with bias and polarised views around the evaluation itself. For example, implementers, donors and the evaluation community may hold differing and contested views on the feasibility, ethics and appropriateness of using a given standard for evaluation designs.

More limited use of common data collection tools: Certain data gathering tools, such as FGDs, may have more limited application in evaluations that look at protection in humanitarian action because of issues relating to stigma and fear of recrimination.

Insufficient access to a representative sample: Factors such as limited time, logistical or security constraints, or even uncertainty about who is affected by protection concerns in the overall population, may limit the extent to which an evaluation team can plan for and achieve data collection that is sufficiently representative to draw generalisable conclusions about target populations. These risks can be mitigated by careful planning to know where people are likely to be and when, and to take into consideration issues of gender and ethnicity that can compromise access to different populations when deciding on the composition of the evaluation team.

Selection of the different primary and secondary sources of data in evaluating protection in humanitarian action should consider their possible advantages and disadvantages and their vulnerability to different types of bias.

Table 3: Data types and sources, their relative strengths and weaknesses and vulnerability to bias

	Examples	Possible use in evaluating protection in humanitarian action	Possible constraints and vulnerability to bias
Primary data generated from individuals	Incident reports, Eye witnesses (including through mobiles), Testimonies, Surveys	Magnitude of violence, Types of incidents, Perceptions of safety and security	Individuals may fear sharing information, even to those deemed 'local', due to fear reprisals or stigma, Quality of data is highly dependent on skills of the interviewers / researcher / or evaluators gathering the information
Secondary data from local stakeholder groups, CSOs, NGOs and other international actors	Focus group discussions, Panels, Surveys, Monitoring reports, Self-reporting / self-assessments reports	Perceptions of safety and security, Understanding and contextualising perspectives, attitudes and behaviour in the affected populations and programme participants and how these change over time, Reconstructing /validating / testing logic models and theories of change, Understanding anomalies / outliers in survey results, Illuminating cases selected in purposeful sampling approaches	Social desirability bias, Group effect bias that may skew results towards uncontroversial and commonly held views, Conscious partiality of data providers
Secondary data from official and administrative sources	Police records, Court records, Hospital records, Morgue records, Demographic and health surveys	Numbers of crimes, Deaths, Violent events, Prevalence studies, Incidence rates of domestic and sexual violence	Often unavailable – inaccessible, infrequent, inconsistent, lack internal validity (i.e. is the same thing being measured over time using the same set of measures), Lack of reporting (due to stigma, recriminations, or discretional use of power by law enforcement officers), Bias from usage: they only capture cases that have been reported or detected or for which services were sought*, Poor state capacity to collect data If the official data set being used has questionable reliability then caution is needed in how it is used – the evaluation report could be quoted out of context, and give extra credibility to an unreliable source data set,

	Examples	Possible use in evaluating protection in humanitarian action	Possible constraints and vulnerability to bias
Secondary data sets from occasional country-specific data sets	Periodic country-specific perception surveys, Ongoing periodic country-specific surveys E.g. Small Arms Survey multi-year 'Sudan Human Security Baseline Assessment, Event monitoring mechanisms E.g. Risk Management Office established by DFID and GTZ in Nepal (now discontinued), Collected data from programme monitoring reports, staff reports and media monitoring	Type of incidents by geographic location in a country, Trends in violence episodes, Perception of safety, Dataset on violent events	Sporadic release and update, Data gaps, Lack internal validity (i.e. is the same thing being measured over time using the same set of measures)

Sources: Compiled and adapted drawing from Hext Consulting (2012) and Church and Rogers (2006: 206-210)

* NOTE on administrative and official records of incidence rates: they are particularly vulnerable to usage bias because the data may show an increase in violence, when actual incidence rates could be decreasing. This could result from an improvement in information systems, from improving levels of to trust in the police / other reporting systems. Thus the data may show increasing levels of violence that previously went unreported.

Special considerations in primary data collection

Primary data collection about protection incidents and overall incident rates is particularly contentious in protection. Evaluators should proceed with great caution before deciding to collect primary data about protection incidents. There are significant concerns with this and evaluators should consider asking why data from incidents rate is needed in the evaluation. Furthermore, incidence rates can be seen as more pertinent to programme design rather than evaluation, so it may be inappropriate or unnecessary to collect such data for evaluation purposes if the programme has collected them already.

Where the risks of re-traumatisation or reprisals are high, there may be serious repercussions for informants. Ethically, it may not be defensible to ask an individual about episodes of violence and trauma if the evaluator does not then link them to a service to counsel and possibly address that trauma. However, it is generally beyond the scope and capacity of the evaluation to put in place the protocols and safeguards needed for this. Engagement with the organisation being evaluated is essential to address this risk.

Box 7: Example of ethical procedures in a pilot study in Rwanda to develop a child protection index to measure the strength of a child protection system

Human subjects research ethics standards require that a caregiver give some form of permission for a child to participate in research, with exceptions only made in extreme circumstances.

In a pilot study in Rwanda to develop a child protection index to measure the strength of a child protection system, data collectors were trained to be aware of the effects that questions may have on the respondent and how best to respond, based on the respondent's level of distress. They were instructed not to provide any counselling, but instead to inform respondents of services available and how to access those services if needed.

The Association of Volunteers in International Service-Rwanda and UNHCR Rwanda agreed to exempt researchers and data collectors from any existing mandatory reporting policies of abuse and violence. When a case was identified, the respondent was informed of services, and asked if s/he would like assistance in accessing those services.

Source: Meier, Muhorakeye, Stark (2015: 35-37)

There are some alternatives to collecting primary data from individuals. For example, incidence rates could come from other sources – such as service providers (as illustrated in Box 8 about GBV data) or from those conducting advocacy on the same issue. Below are some suggestions for the measures that can be considered when data collection touches on protection incidents:

- Engage those actually providing a relevant service to conduct the data collection, as they are able to link the informants to the service
- Rather than asking people to recount their own experiences, ask about someone else who has had this sort of experience (mother, sister, etc.)
- Consider using interviewees with some basic counselling skills
- At a minimum, ensure that interviewers have sound interview skills.

Box 8: Challenges and constraints in gathering and using GBV data

GBV is difficult to quantify as many cases go unreported, its scope is difficult to estimate and existing data is often misunderstood, misrepresented and ineffectively utilised. … Prevalence studies can provide some idea of the overall picture of GBV in a country or area. However, they are only estimates and generally provide little information more subtle or short-term changes in GBV trends.

Source: Bain and Guimond (2014: 16-17)

Box 9: Conducting surveys in access- and data-constrained environments in conflict and post-conflict settings

In north-east Afghanistan a survey was conducted in a situation where official population data was not available. In order to develop the sampling frame, interviews were held with the village council (Shura) to determine the number of households in the village, and then this data was used to calculate the number of interviews to conduct in that village. Some areas had no maps at all, and not even agreed names for villages, making it very time consuming to determine a sampling strategy.

The same survey commissioned quarterly reports on communities and districts in which the survey was conducted to capture significant local events significant changes, disasters etc. which helped identify contextual factors which were key in analysing the survey findings.

Random sampling emerged as the best practice in north-east Afghanistan:

> 'We opted to collect a random sample of households in every community, for every survey. An alternative would have been to collect panel data – that is to sample the same households for both surveys. We did not do this because we were afraid of high attrition. We anticipated that a deteriorating security situation would have forced many households to flee or be on the move for work. We also wanted to minimise risks for our respondents, households who speak too often to foreigners might have been at higher risk of reprisal by insurgents, which could in turn affect our responses.' (Böhnke, Koehler and Zürcher, 2014: 112)

In Nepal the success of the Chitwan Valley Family Study panel surveys during a period of armed conflict was due to flexing the process around the context, notably:

- Interviewer calling times were altered to only be conducted during daylight (to reduce suspicion)
- In peak violence periods, the data collection was switched from monthly to quarterly, and some locations were stopped entirely
- Respondent tracking was expanded to include institutionalised populations (military / in prison).

The survey team found these measures enabled a strong response rate, and thus a strong survey quality:

> 'Our results are also consistent with the conclusion that different dimensions of armed conflict can affect survey data quality in opposing direction, with higher numbers of bombings in the local area slightly increasing refusal rates but higher number of nationwide political events actually increasing response rates through decreasing both refusals and non-contacts.' (Axinn, Ghimire, and Williams, 2011: 23)

Consider alternatives to primary data collection from individuals

For the reasons discussed above, the need to look for alternatives to primary data collection from individuals is particularly high in evaluating protection in humanitarian action. Data about incidents can also be extrapolated through exploring proxy indicators.

Data sources beyond interviews with individuals are often overlooked. However, there is some untapped potential for drawing on official data, which is a growing area in monitoring and evaluating conflict crime and violence programming (CCVR, 2012).

Some humanitarian responses are occurring in countries with reasonable state capacity, such as Indonesia, Philippines and Lebanon, in which official data may already be of a good quality. Most notably there may also be ongoing donor investments to improve state capacity in official data in areas of relevance to protection in countries with weaker state capacity.

Consider gathering and using primary data from service providers

Gathering and using primary data from service providers is seen as good practice in protection-related programming as well as in evaluation.

For example, a service provider that classifies stages of healing following trauma could report on how those using their services are progressing through the different stages. However, service providers sometimes lack capacity for data collection, and data collection for specific evaluation purposes may be a low priority. Ideally, the intervention being evaluated may have capacity development components to enhance service providers' monitoring capacities, but this is seldom the case in humanitarian interventions.

Evaluators (and especially those responsible for strengthening monitoring systems) may therefore need to consider including some elements of capacity strengthening whenever service providers are expected to take part in data collection work specifically commissioned for an evaluation. There is also a risk of bias, particularly if the service provider is directly supported as part of the programme being evaluated. There may be also options in terms of peer-to-peer data collection, as described in the following example from Search for Common Ground.

EXAMPLE: Search for Common Ground

In DRC, Search for Common Ground partnered with and supported local women's groups to provide services for women, and also trained them in gathering data for M&E purposes. That appeared to have enabled women to talk more openly about taboo subject matters as it was a case of local women talking to local women and being then linked to locally available services.

Evaluators may invest in an analysis of the context and the interests of different stakeholders to help mitigate this bias. Even literature reviews – as suggested in the evaluators' insight below – can be helpful in this respect. See "TOOLKIT item #5 – ODI/RAPID influence and interest matrix" on page 95

EVALUATOR'S INSIGHT on dealing with bias

In order to cope with the expected bias from informants, the multi-donor evaluation of conflict prevention and peacebuilding programming in South Sudan used the field based interviews to verify the literature review rather than using them as the key data source:

> Importantly, the evaluation was not dependent on these field level interviews and discussion groups - which might be seen as partial or biased – but rather these were used to triangulate the more substantial evidence and preliminary findings from the first stage literature review and analysis. (Barnett and Bennett, 2014: 45)

SECTION 12: Data on the less tangible and harder-to-measure dimensions of protection in humanitarian action

Data on perceptions

Many in the conflict prevention and peacebuilding field are very confident in using perception-based data – and many would claim that 'perception is as important as reality'. Individuals are often motivated to commit violence on the basis of a perception – how they view another group, or rumours about tragedies – and indeed much peacebuilding work is about trying to change such perceptions. Some evaluation users, however, may be sceptical and may see data regarding perceptions to be a weak proxy for 'hard data' on actual violence, displacement or other variables.

The security and justice field also works with perceptions – particularly perceptions of safety and security. It can be more politically and technically viable to collect perception measures of safety and security (such as the percentage of men and women who fear a crime) than incidence data (number of incidents of violent crime). Examples of data-collection tools focused on perceptions that could be useful here include:

- Movement maps that visually capture where people feel safe to move, possibly mapped over time to show changes in perceptions of safety and security.
- Body images, where women are invited to talk about gender-based violence through drawing bodies and describing what they are most proud of and least proud of in their own bodies.

It should be noted that such tools are highly specialised and using them effectively and sensitively require a significant investment in training and engagement of appropriate team members.

Perception-based data should not be used in isolation. Three considerations apply here:

- It should be triangulated and analysed together with other sources of evidence about the programme and/or the context (for example media reports of violent events can be a good corroborating indicator).
- The utility of perception-based data depends on the degree to which changed perceptions (for example, of gender roles and violence) were part of the programme objectives.
- Perceptions and incidence data may not align: for instance, there could be a time lag between an actual improvement in crime statistics and perceptions reflecting this. There may also be significant variations across short geographies: a village that suffered an atrocity will have a very different sense of security that one nearby that did not.

It should be kept in mind that, by contrast, in other fields of humanitarian action many evaluation stakeholders have less confidence in measuring perceptions. Actors accustomed to relying on 'hard data' on malnutrition, litres of water available per person or disease vectors may be inherently sceptical of data regarding perceptions.

Box 10: Measuring community security – Saferworld

Saferworld implements a large spread of community security programming globally. It promotes a participatory approach to monitoring and evaluation, with the communities themselves determining the dimensions to be measured. Saferworld's Community Security Approach addresses insecurity at three levels and conducts measurement at all of them: community/local level, sub-national/district level, and national level. The measurement of community security combines both perception-based data (e.g. sense of security) with more tangible dimensions of security (e.g. number of attacks on community). These measures span all three levels.

Community / local level

As part of the programme, the community identify security concerns and generate action plans, and determine how to monitor progress and measure success. Key dimensions to change that should be measured at this level include:

- Specific outputs to be achieved (e.g. establishment of local police post)
- Changes in the way the community feels about itself and agency (e.g. willingness to tackle sensitive areas of concern)
- Changes in the perception / sense of the community about their situation (e.g. do they feel safer? Do they trust their authorities?)
- Changes in relationships within the community and/or with others (e.g. relationships with the police service)
- Changes in the behaviour of the communities and the security provider.

Sub-national/district and national level

The programme uses research and advocacy to link local improvements up to sub-national and national levels to promote policy change that draws on the local programme experience. Key dimensions of change that could be measured at this level include:

- Behaviour of security providers towards communities (e.g. number of attacks by security providers on individuals and/or communities, the extent to which security providers see themselves as a service to the community, rather than a force for control)
- Community behaviour towards security providers (e.g. willingness to report crime or security issues to relevant authorities)
- Relationships between the community and security providers (e.g. quality of interaction between security providers and communities in meetings, level of continued reliance by communities on non-state, informal security providers)
- Feelings of safety and security (e.g. the proportion of women who feel confident of walking in the community after dark)
- Changes in the way sub-national and national security providers consult, engage and respond to communities
- Changes in how security budgets are defined and used.

The programme uses participatory evaluation process in which the community convenes for a day to identify transformations in the relationships and behaviours behind insecurity, and how these changes have affected experiences of security. Annual community security assessments have been a valuable data collection instrument for the programme who identify the nature of security in that locality and can track specific security issues, the availability of services, and the feelings of safety of the communities. Source: Saferworld (2014)

Sense of security

The sense of security is a particularly important aspect of judging outcomes related to the environment for protection and the perceived relevance of interventions, and may even help understand the sustainability of the changes it induced.

Security indicators need to be developed early in the programme so they can be used in monitoring. Without a baseline and monitoring data, it is likely to be too late to make meaningful measurements in a summative evaluation. Indicators are best developed as 'participant-generated indicators' by asking people in the community about what they believe indicates that the situation is safe. For example, in Darfur IDPs stated that they would send 1-2 members of the family back and wait to see if they could remain safely in their location of origin throughout one agricultural season.

All such participant-generated indicators should be disaggregated by age, gender, and any other salient distinctions.

A community security assessment or focus group discussions in same-sex groups can be used. A basic question would be 'under what conditions would you feel safe doing xxx activity' (collecting firewood or income generation or moving between location A and B etc.).

Such data can contribute to a formative evaluation, providing insight into the relevance of the intervention.

Box 11: A sample of perception-based indicators

An Itad report assessing a suggested list of governance and conflict indicators on behalf of DFID endorses several perception-based indicators (for use in conjunction with objective indicators), some of which are relevant to protection:

- Percentage of citizens who say they feel safe going out in their neighbourhood at night (disaggregated)
- Percentage of citizens who believe bribes are necessary to access police services
- Percentage of target population who report positive attitudes to civil-military relationships and to reintegrated combatants
- Percentage of community who do/don't feel threatened by presence of ex-combatants.

Source: Barnett, Barr, Duff and Hext (2011)

MODULE C – ANALYSIS

» This module is primarily addressed to evaluation teams, but it is also relevant to evaluation offices, particularly in assessing the quality of evaluation reports.

Content of this module at a glance		
	What is this module on?	Who is this module for?
SECTION 13	Suggests how evaluation teams should revisit the original intervention logic as a point of departure for their analyses	Primarily for evaluation teams
SECTION 14	Reviews the concepts of causality, attribution and contribution and how they are likely to be applied in EHA protection	Primarily for evaluation teams but also useful for evaluation commissioners when assessing the quality of evaluation reports
SECTION 15	Presents insights from other fields that are of relevance for analysing influence on the protection environment	Primarily for evaluation teams

SECTION 13: The starting point for analysis: Revisiting the intervention logic

See "TOOLKIT item #4 – ODI/RAPID approach to Theories of Change" on page 94 In evaluating protection in humanitarian action, challenges are frequently faced when it comes to expanding on the critical reflection on the intervention logic or theory of change that began when the evaluation was initiated. Uncertainties about the intervention's approaches need to again be unpacked and reconsidered.

We start this module with a discussion of a number of fundamental factors that may skew analyses that evaluators should be aware of:

- **Weak designs and under-developed intervention logic**
  Projects and programmes developed hastily in response to an acute crisis may lack a theory of change or other intervention logic that articulates how activities and outputs were expected to address protection needs, much less the assumptions about how the initiative was expected to overcome obstacles inherent to the conflict context. An evaluator is sometimes tasked with providing guidance for a revised or enhanced theory of change for future programming.
- **Influence of institutional worldviews**
  Analysis of the drivers of protection risks and needs in a given context may be driven more by institutional worldviews or prevailing policy narratives rather than deep contextual knowledge, thus skewing the assumptions underpinning programming. An agency may ignore potential mismatches between programming and protection needs if they have operated under a 'if all you have is a hammer, everything looks like a nail' approach to programme design: providing the goods and services they normally supply, rather than those required by the situation. This

may sometimes limit the parameters of the evaluation as well when terms of reference fail to provide room for questioning the unconscious worldviews that frame programming. This can even lead to a narrow evaluative focus on whether the intervention is 'doing things right' (within standard agency modalities) when there are major unresolved questions regarding whether the intervention is 'doing the right thing' (in terms of protection needs). For this reason an evaluation with a narrow focus on effectiveness may have an entirely different conclusion to one focused on broader relevance.

- **Complexity and a focus on 'doing'**
  If there are multiple components or different protection actions nested within a larger intervention, there may be a lack of detail on what is actually expected to be implemented for whom, by when, where and how. As a result, there may be undue influences or deep-seated bias affecting what gets implemented, assessed, measured and monitored.
- **Perceptions and more hidden dimensions of results**
  A sense of safety or security is much harder to describe and measure as they it has more 'hidden' components than other, arguably more tangible areas of results linked to assistance provided in terms of health, nutrition or sanitation. Donor demands for concrete evidence of 'results' can sometimes aggravate these limitations.

For these reasons, among others, evaluators should be aware of some key questions:

» Be alert to the **different ways of talking about protection** that agencies use to describe and frame. Note that some agencies may use similar terminology to mean different things.

» Be aware that there may be some disconnect in how different teams and programmes within the same organisation – and even within the same operation – see themselves working within and around protection, including their assumptions, and how this may influence the logic of the intervention. This is particularly important in interventions where **protective** actions are being implemented by staff from other sectors.

» Be aware that different **ways of approaching** protection in humanitarian action are likely to co-exist in the same programme, intervention or context. This is likely to complicate analysis because it affects the extent to which evaluators will be able to identify and 'isolate' t**he specific elements/factors contributing to protection outcomes of interest.**

» Be alert to how the international legal frameworks applicable to the different contexts and specific situations in which humanitarian actors operate (e.g. in international and non-international armed conflicts and other situations of violence) can affect the relevance, feasibility, connectedness, coherence and effectiveness of certain types of protection actions.

By unpacking the often diverse implicit and explicit expectations and assumptions across the results framework, analysis of intervention logics and theories of change can also help focus lessons and recommendations in ways that resonate with different users' decisions about whether, where and how the intervention could be scaled up or carried out in other settings under specific conditions

(Bamberger, Rugh and Mabry, 2012: 183; 227). Even if there are dangers that a narrow focus on certain 'results' narrows the perspectives of certain users, analyses of such results can be used as entry points for a broader discussion about the protection-related issues that impinge on those targeted actions.

Specific uses of programme theory in evaluation that can boost the explanatory strength[23] of an evaluation include drawing conclusions regarding:

- The strength and weakness of the intervention logic underlying the design of the intervention
- The strength and weaknesses of how the intervention was implemented
- How contextual factors contributed to, or militated against, the achievement of intended results, thereby raising attention to the contextual relevance of the intervention modalities
- How the intervention affects, and is affected by, different groups, for example the extent to which 'do no harm' principles were applied and due attention given to gender perspectives and issues related to marginalisation and vulnerability (Bamberger, Rugh, Mabry, 2012: 182)

Critically reflecting back on programme theory can also be important when considering whether there may be alternative explanations for the changes in, for example, the perceptions of safety in a community or the actions of authorities.

Three examples of intervention logics for protection

To provide a better sense of entry points for analysing different types of programmes, this guide presents three illustrative examples of how protection can relate to a programme. They are intended as a tool for reflection and do not suggest a typology of protection programming or point towards a set of good practice.

The examples illustrate different ways in which humanitarian actors address and weave protection into interventions. The order of the three examples is in no way intended to suggest judgements on their relative appropriateness. The examples are used to show a range of entry points for protection among different agencies.

23 This is particularly relevant when using change-centred approaches to evaluation.

Box 12: EXAMPLE 1- A programme where protection is to be achieved through specialised or dedicated actions

General features

Evaluators may be asked to evaluate a project or programme where protection is achieved through specialised or dedicated actions.

These programmes are often described in the literature as vertical or stand-alone protection actions.

They are often characterised as the traditional remit of protection actors with a specific mandate anchored in international legal instruments and of other actors (including NGOs and INGOs) with specific expertise in thematic areas – forced displacement, child protection, or working with people with disabilities, the elderly and sexual and gender minorities.

Programme features from a protection angle

The desired outcomes of these interventions explicitly articulate and speak to protection issues (Davies and Ngendakuriyo, 2009; de Sas Kropiwnicki, 2012). There are expectations that protection actions in this example are:

- Informed by a protection-specific mandate, policy and or strategy, and that the analysis is explicitly used to inform the protection strategies and logic of the intervention/programme
- Explored in the conflict analysis that should generally underpin and inform humanitarian response strategies
- Anchored in different modalities and lines of work connected to protection (as discussed in the introduction section of the guide).

Murray and Landry (2013: 5) note that protection actions in this example are usually featured in the 'protection chapter' of some key humanitarian funding tools such as the consolidated appeal process.

Some protection actions of this type may aim at influencing outcome-level changes in the broader protection environment.

Protection actions such as these are likely to be highly sensitive to agencies' mandates. There are greater expectations that agencies with a specific protection mandate have greater capabilities to raise fund, design, carry out, lead and coordinate interventions and dedicated protection-oriented programmes and services of this type.

Box 13: EXAMPLE 2: A programme where protection is integrated into other sectoral and multi-sectoral interventions

General features

Alternatively, an evaluator might be asked to consider a project or programme where protection goals are to be achieved or supported by integrating protection-oriented activities into other sectoral and multi-sectoral programming. This implies applying protection-related perspectives and activities within an intervention such as water and sanitation, education, health, food security, livelihoods, or shelter.

Expected programme features from a protection angle

Protection actions in this example are likely not to be explicitly anchored in a given agency's mandate. Services and actions oriented towards addressing specific protection needs or reducing exposure to protection risks are integrated into other ongoing sectoral and multi-sectoral interventions. As a result, protection-related objectives may be vaguely formulated and/or based on a relatively weak analysis of the overall protection context. On the other hand, a sectoral perspective can also reveal new protection needs and challenges that are not apparent to 'protection experts' whose frames of reference are more focused on 'conventional' protection concerns.

Protection actions such as these are also likely to encompass a variety of implementing approaches – from information provision to operational advocacy and provision of specialist services.

Box 14: EXAMPLE 3: A programme without an explicit 'protection' component, but which should be expected to fulfil expectations to 'do no harm' or, more generally constitute safe, accessible and/or dignified programming

General features

Many evaluators encounter programmes which, while not explicitly incorporating protection objectives, benefit from the evaluator using a protection lens (or tools and approaches related to the evaluation of protection). These programmes will often be designed according to, and may make reference to, various principles and frameworks mainly relating to safe programming,* safe access to assistance and services. Some evaluations may be tasked with exploring whether the intervention should have applied a protection lens, for example by applying a 'do no harm' perspective (see Box 5).

Some programmes, rather than actively promoting protection objectives, may be designed to avoid causing further harm: this often requires iterative adaptation and mid-course corrections based on changes in the safety of those accessing (sectoral) services or the delivery of assistance in different sectors. For example, the location of assistance distribution points or the timing of service delivery can be adapted based on changes that can enhance the safety and access to services.

Expected programme features from a protection angle

Some of the main points arising from this example are:

- Protection actions are not always a matter of protection-specific mandates and they are not necessarily undertaken by protection specialists
- Programme documents might not include any direct reference to 'protection', even though they may involve significant protection goals
- There may be references to agency guidelines, agency-specific commitments relating to protection in humanitarian action, but not to how these should be translated into programming actions – including in resource allocation – in specific sectors.
- Protection may even be overlooked entirely in the intervention logic, which in turn is likely to place the onus of reconstructing the protection-specific intervention logic on the evaluation team.

* NOTE on the term: 'safe programming': Safe programming refers to any attribute and way of work of the programme/service or other type of intervention that aims to ensure that: (a) interventions do not put the population in danger; (b) interventions contribute to their security as much as possible; and (c) potential threats are analysed and monitored in a systematic way.

Source: Oxfam GB (2009)

SECTION 14: Analysing causality, attribution and contribution

Causality

Analyses of causality need to start with describing the baseline, which could include existing levels of service provision, processes already under way, the current situation of human rights abuses, the institutional setting, and so on. Ideally this will have been done as part of planning the intervention, but in humanitarian settings this step is regrettably often forgotten or undertaken in a rudimentary manner, which can create an additional burden on evaluation teams who must then 'reconstruct' the baseline by drawing on a range of secondary data about the situation at the start of the intervention, or stakeholder recollections.

Analyses also need a description of the protective actions being taken, including inputs, activities and outputs. This may also include description of internal constraints (e.g., budget, human resources, etc.) and external factors and events impinging on the intervention.

The next step is to assess the relational assumptions in the programme's explicit or implicit intervention logic. This involves establishing whether a relationship between two or more phenomena is assumed to exist and, if so, its direction and magnitude. The empirical data gathered in the evaluation itself may significantly change the evaluation team's understanding of who holds these assumptions and how they are interpreted in practice.

Most evaluations are steered by normative analyses – that is, they compare 'what is' with 'what should be' and the current situation with a specific target, goal or benchmark. However, many humanitarian interventions are focused heavily on delivering a set of outputs, with grander normative objectives described in somewhat vague or visionary terms. This means that the evaluation team may also need to reconstruct a more realistic theory of change based on either (a) the intervention logic as perceived by key stakeholders in the programme, or (b) the evaluation team's or commissioning officers' own assessment of what would constitute a more plausible theory of change.

Analysing quality and value – Evaluations almost invariably ask about the overall conclusion as to whether an intervention can be considered a success, an improvement compared to the previous situation, or the best option (Rogers, 2014: 10). Some authors refer to this as asking **truly evaluative questions** (Davidson, 2004: xii) to underscore that what makes evaluation different from other endeavours (e.g. performance measurement and monitoring) is asking '**how good**' and '**how valuable**' the results of a certain intervention are for specific groups and individuals and why. In this sense, truly evaluative questions are those that do not stop at asking 'how things have changed' and 'to what extent has the change been brought about the intervention being evaluated' but also **examine the importance, quality and value dimensions of change**.

Causation (or causal) analyses seek to **establish the intervention's role in producing the results described or implied in the (reconstructed) intervention logic**. One central concern when answering causal questions is documenting that a given result, change or effect has been caused by the intervention and not by coincidence or by other concurrent factors at play. Particularly in the complex and dynamic contexts that characterise humanitarian emergencies, it is essential that the evaluation does not assume that **correlation** (e.g., a change in the frequency of protection violations) can be equated with **causation**. Contribution analysis (discussed below) is a way of unpacking the question of relations between interventions and actual phenomena.

Special considerations apply when analysing causality in programmes and interventions in which protection is more implicit or has been woven into other (non-protection-oriented) services, activities and programmes.

Establishing causality is likely to be a more complex and resource-heavy exercise because of the work needed to identify and reconstruct the 'bundle' within which protective elements have been infused.

Attribution and contribution

Establishing causality is not straightforward, particularly in crisis, fragile, conflict and post-conflict settings, and there are different ways of looking at causation (see Box 15). One of the pervasive challenges with establishing causation in evaluation is that it may not be possible to isolate the results brought about by a given intervention amongst a host of other contextual factors. This point is commonly referred to in evaluation as the **attribution problem**.

Attribution requires establishing the causal implications of an intervention and/or the causation of an observed phenomenon (Scriven, 2010: I; also see Gerring, 2012). However, especially in the context of evaluating humanitarian action, it is rare that causal attribution refers to **sole attribution**. Rather, it often refers to establishing **partial attribution** or **analysing contribution to impacts**[24]. See "TOOLKIT item #6 – Overview of contribution analysis" on page 97

A range of techniques exists to help evaluators examining causality – whether in terms of sole attribution, partial attribution, or contribution to results[25]. At the broadest level, analysis and techniques used to answer causal questions in evaluation will pursue one or more of the following.

Factual analysis involves asking: what kind of results and changes (outcomes or impacts) occurred for whom in a given context? How did actual results of the programme or intervention compare to those expected from the logic model or theory of change that informed the intervention? Are the results and changes that can be observed consistent with the theory?

24 This same point is echoed in the context of evaluating peacebuilding, conflict transformation and aid in conflict settings. See for example Chigas, Church, Corlazzoli (2014); Church and Rogers (2006); Andersen, Bull, and Kennedy-Chouane (2014); Scharbatke-Church (2011).

25 This brief summary draws from Chigas, Church, Corlazzoli (2014: 20)

Analysis of alternative explanations involves examining different scenarios posing alternative explanations (other than those related to the outputs of the intervention) that could account for the observed changes and results.

Counterfactual analysis produces some estimates or seeks to explain what would have happened if the intervention had not occurred. Conventional attribution analysis requires the group receiving the programme or intervention to be matched to a comparison group. Here, there are stringent requirements for dealing with bias (e.g. selection and contamination bias)[26] and for dealing with alternative explanations of the observed changes, which ideally should be eliminated. While common in many forms of evaluation, rigorous counterfactual analysis is rare in evaluation of humanitarian action.

Contribution analysis seeks to assess the extent of the influence of a particular actor in contributing to the overall changes resulting from the collaborative technical and financial intervention carried out by other actors in a given context. (Bamberger, Rugh, Mabry, 2012: 404). Contribution analysis (Mayne, 2001) is also the label that indicates a specific technique used to establish contribution in a structured manner following six steps:

1. Develop the theory of change
2. Assess the existing evidence on results
3. Assess alternative explanations
4. Assemble the performance story
5. Seek additional evidence
6. Revise and strengthen the performance story

26 Biases include selection bias (i.e. areas receiving humanitarian assistance are likely to have attributes that make them more or less likely to recover, compared to the average), and contamination bias (areas targeted by one actor are also likely to have other sources of assistance that may make it difficult to separate the different sources of changes) (Puri et al., 2014: v).

EVALUATOR'S INSIGHT with a word of caution on using programme theories to assess causation

Programme theories (or logic models) in evaluation have often been misused and affected by the poor quality of the process through which the programme theory was developed in the first place, and by the poor quality of the end product. There is growing scepticism about the term, which has been associated with use of new jargon to refer to old concepts and pressures to insert new reporting demands (theories of change have been criticised as being 'logframes on steroids').

Indeed there are an increasing number of instances where the theories are overly donor-driven; where the process to develop them has been extractive and non context-aware; where ownership of the tool by programme staff is so low that once the programme theory is developed it is never looked at during the course of the intervention. These dysfunctions have often been particularly associated with rigid or top-down use of results framework approaches.

A well-designed programme theory (or logic model) can greatly strengthen an evaluation design and support the analysis of findings. However, it is controversial – and still widely debated in evaluation theory and practice – whether a well-articulated programme theory model can also help test causality.

Some would argue that if implementation proceeds according to the theory and if the expected outcomes are achieved, this gives some basis and credibility to the claim that the programme contributed to the outcomes. Even in those evaluation scenarios however, the contribution's claims would be more robust and credible if alternative models (alternative theories for how change is expected to happen) are developed to test rival hypotheses.

Source: Bamberger, Rugh and Mabry (2012: 484)

SECTION 15: Insights from evaluating advocacy and other initiatives intended to influence the protection environment

Evaluating advocacy[27] in aid and development settings is a growing area of practice within the broader practice of evaluating advocacy, policy influence, communication and campaigning.[30] With few exceptions, there is a dearth of humanitarian-evaluation specific guidance looking at advocacy.[31]

Many of the challenges associated with evaluating advocacy efforts can be found in other contexts, but several are more prominent in humanitarian contexts. Recent UNICEF guidance on monitoring and evaluation of advocacy describes those challenges as:

- **The speed of decision-making and the urgency of information needs:** During and after an emergency, a quick systematic assessment is necessary to inform decisions being made about advocacy efforts. However, the nature of emergency and post-emergency settings often impedes a quick systematic assessment of the conditions being conducted.
- **Inherent volatility and complexity:** Due to the volatility and complexity of emergency and post-emergency settings, identifying advocacy targets can be difficult. 'This poses difficulties not only in conducting advocacy in the first instance – and hence in demonstrating its effects in light of a rapidly changing landscape – but also in accessing the most qualified stakeholders who can shed light to the evaluation team on UNICEF's efforts' (Coffman, 2010: 14).
- **The abstractness of advocacy processes can make data collection difficult:** Advocacy processes also have abstract outcomes that are difficult to define precisely (public will or political will, for example). As such, less conventional methods are applicable to advocacy efforts (Coffman, 2010: 20).

27 For the purpose of this guide, advocacy is defined as any types of action or intervention that requires some form of influencing work (Tsui, Hearn, and Young, 2014:11).

28 For an overview see for instance Stachowiak (2013), Tsui, Hearn and Young (2014), LFA (2013).

29 In 2010 UNICEF completed a sizeable attempt at documenting, systematising and producing guidance on monitoring and evaluating advocacy including a specific section on advocacy in the context of crisis and emergency and post emergency contexts. (Coffman, 2010).

Box 15: Comparable challenges in evaluating advocacy and influencing work

Evaluators of protection in humanitarian action could benefit from looking at the practice of evaluating advocacy and policy influence in the broader aid and development settings. The two domains grapple with a comparable set of evaluation and measurement challenges, including:

Causal relationships: Linking advocacy and outcomes is complex.

Subjective gains: Defining success is challenging and varies depending on who is asked. The goal posts can often shift depending on the circumstances.

Multiple approaches: Influencing policies and influencing behaviour change can be part of many approaches including lobbying, advocacy, policy research or campaigning. It may be difficult to assess which approach leads to which results at outcome and impact level.

Programme approaches are inherently more speculative than direct interventions, and the benefits are less easily articulated, typically less quickly achieved, and also less easily assessed or measured. Long horizons: Advocacy and influencing work are long term. Change can be slow and incremental.

Conflicting political process: Influencing often means engaging in a process that may have political consequences, which in crisis and conflict situations may be even more far-reaching and draw the evaluation into sensitive and contested areas related to humanitarian neutrality and impartiality.

Tension about metrics: There is a tension between the desire for 'metrics' or quantifiable indicators and the need for usefulness analysis of progress. Many metrics are either too narrow or short term, focusing on activities such as the number of newspaper citations, or too broad or distant, for example changes in policy or legislation.

Focus on **measurable data** in advocacy evaluation (e.g. data from social media, news stories, etc.) tends to be far away from the real value, far away from actual change, and so comparatively uninteresting for users when there are demands for evidence that advocacy and campaigning is delivering tangible results.

Most outcomes and impacts are hard to see: In value terms, advocacy and campaigning is an iceberg: most of the impact may be submerged and hard to see. And the temptation to focus only on the part that is visible risks creating a radically false picture that generates misleading information and so encourages poor decision-making.

Sources: Coe and Majot (2013); Reisman et al. (2007); Chapman and Wameyo (2001); Tsui, Hearn, and Young (2014); Schlangen and Coe (2014)

Below is a compilation of insights into evaluating advocacy and other initiatives intended to influence policy. The intention is to highlight some of the emerging learning and guidance generated in that field that may resonate with those evaluating protection in humanitarian action.

Move towards a 'try and evolve' approach to monitoring and evaluation

Snowden and Boone (2007) have suggested that the appropriate management style for complex interventions such as the case in many advocacy actions is to use an experimental 'try and evolve' approach, which recognises that even successful interventions will involve missteps or mini-failures. In that light, identifying and learning from these missteps is essential to guide programming and should not be understood as lack of effective planning and design – as it is the case in some 'conventional' monitoring and evaluation guidance (Tsui, Hearn and Young, 2014).

Focus on evaluating progress and contribution

Over time, advocacy strategies evolve. As a consequence, activities and desired outcomes also change. Course-correction and adjustments are the most realistic expectation in monitoring and evaluating advocacy. That is one of the reasons why evaluating progress is also important. 'Advocacy M&E typically focuses on the advocacy journey rather than just the destination' (Coffman, 2010: 2). That journey has usually started before the intervention and will continue long afterwards. Evaluations should recognise this broader perspective at the outset.

In advocacy evaluation, there is a strong focus on articulating and measuring interim outcomes because ultimate goals (passing a resolution, or changing an entire policy approach) can have very long time horizons – years or even decades. An important focus of advocacy evaluation, therefore, is interim outcomes, which (LFA, 2013: 5):

- Are benchmarks or milestones that demonstrate incremental progress toward your ultimate goal (e.g. getting an important policy-maker on board as a champion)
- Can be the direct outcomes of your advocacy activities or tactics (e.g. after meeting with an important policymaker, they commit to authoring a bill)
- Are often outcomes that you must achieve in order to reach your ultimate goal (e.g. you need a certain set of policy-makers on board in order to get a bill or a resolution passed).

Distil evaluative information meaningfully

Distilling information down to the basics is an appealing, efficient and necessary way to communicate what happened – particularly if it can be done with numbers. But numbers, instead of being an aid to strategic decision making, risk being a substitute for it. As a rule of thumb, the more complex the context being assessed, the less credible meaning is to be found in a simplified distillation of it. So for organisations trying to assess the value of advocacy and campaigning, translating qualitative information into numbers can devalue this information by stripping it of the

very detail that gives it value. It also typically conveys a false sense of precision and objectivity. For this reason, one working principle in reporting advocacy and campaigning should be 'no narrative-free data' (Schlangen and Coe, 2014: 7), a principle that is highly relevant for evaluating protection in humanitarian action.

Use single and multiple case studies

Use single or multiple case studies. Case studies often examine different aspects of an advocacy effort and collect data from a wide range of stakeholders (those involved in the advocacy effort and those who are its targets). Case studies provide a full and detailed account about what happened. Isolating data points can disguise the full story or context. Multiple case studies are useful when advocacy efforts are based in multiple locations or contexts. This design allows comparisons across cases, which can help in identifying patterns or existing and emerging themes (Coffman, 2010).

Next Steps

The field of humanitarian evaluation – and of the evaluation of protection, in particular – is rapidly evolving. This guide represents an attempt to identify key issues and collate good practice with regard to the evaluation of protection as it exists at the time of writing. However, the guide is by no means exhaustive, and there will be challenges and experiences – both current and emerging – which have not been captured here.

Over 2016 and into 2017, the ALNAP Secretariat aims to pilot this guide with ALNAP Members and other organisations concerned with protection in humanitarian action.

For more details on the pilot go to: http://www.alnap.org/evaluating-protection.

If you, or your organisation, are interested in participating in this pilot, please email us at eha@alnap.org.

Toolkits

TOOLKIT item #1 – Evaluability checklists for evaluation of protection in humanitarian action

This guidance suggests that evaluability studies should cover four main areas[30]:

1. **Overall level of ambition and type of questions** that evaluation stakeholders and programme stakeholders would like the evaluation to answer.
2. **Programme design and intervention logic** – particularly important for outcome and impact evaluations that make use of theory-based designs to understand causation, mixed-methods designs, and outcome-based approaches that look at contribution to results in multi-actor or networked interventions (e.g. outcome mapping, outcome harvesting).
3. **Availability of data** – or the possibility of generating additional data – required for the evaluation to answer the specific questions that commissioners and stakeholders have.
4. **Conduciveness of the context** to carry out an evaluation that looks at protection. This should include considerations around organisational 'climate' and leadership support to the evaluation, considerations of access, logistics, and safety of the evaluation team, and considerations of ethical appropriateness.

Below is a set of four checklists covering those areas.

» **Why use checklists to examine the evaluability status of a programme or intervention?**

Checklists are a very flexible means of ensuring the systematic consideration of a set of issues across all components a programme or intervention being examined. Evaluability checklists can be used as stand-alone tool (i.e. when carry out fully-fledged Evaluability Studies) or they can be incorporated as part of the inception phase of an evaluation, or during the preparatory work needed to develop and refine an evaluation matrix.

EVALUBILITY CHECKLIST 1: Level of ambition and type of questions that evaluation stakeholders and programme stakeholders would like to see answered

In the pre-evaluation stage, or during an evaluability study, there should be opportunities to get a sense and clarify the expectations that different programme stakeholders and evaluation stakeholders have for the evaluation exercise. This can be grasped by asking questions such as the following:

- What are the sorts of question that programme stakeholders would like to see answered in an evaluative exercise or other reflective exercise? Are those questions relating to process and normative aspects of the intervention? Are those questions relating to cause-and-effect issues?
- Is there an expectation that the evaluation will focus mainly on issues and processes internal to the programme or intervention over which the agency should have more control and influence?
- Or conversely, the evaluation is expected to look at a higher level of results (outcomes and impacts) that are beyond the sphere of control and (perhaps) influence of the intervention or programme and even of the agency itself, and that may touch on protection environment-wide issues and dynamics?

30 The elements proposed could be used in general EHA work. They have been modified for this guidance, but they largely draw from, and are in line with EA guidance developed and piloted by different donors (e.g. DFID, NORAD, Sida, USAID) and operational agencies including UNFPA, UNICEF and UN Women.

- Is the evaluation expected to cover process issues and results within the domain of humanitarian action, or to reach across multiple domains in the protection architecture, including development aspects (and global or country level – depending on the questions asked and the unit of analysis)?
- Is there broad alignment (or are there frictions between) the questions that programme stakeholders would like the evaluation to answer, and the questions that funders and other actors external to the programme would like see answered in the evaluation? What are the implications for ensuring broad utility and for the accountability of the evaluation team itself?
- Are there realistic expectations for the evaluation to look at a higher level of results and environment-wide issues in light of the level of resources, time, evaluation expertise, and protection know-how in the team that should carry out such exercise?

EVALUBILITY CHECKLIST 2: Intervention logic / programme design

A protection-oriented evaluability study should help uncover whether protection – with any of its related domains of work and themes – had been incorporated in the earlier stages of the programme life cycle (assessment, design, implementation and monitoring).

If it was, it should be possible for the evaluators to discern how and to what extent protection issues had been spelt out in the protection analysis (if one had been carried out), in the result frameworks used by the programme, and in related monitoring and reporting tools[31].

On the other hand, if protection in humanitarian action had not been incorporated in the earlier stages of the programme's cycle, this would translate into a need for greater efforts when gathering information to support the evaluative judgments (Faúndez and Weinstein, 2014: 11). The timing and scope of the evaluation should also be revised in light of those considerations.

During a pre-evaluation process, or during an evaluability assessment exercise, the type of questions that can be asked can be formulated along those lines:

- Does the programme clearly define the problem that it aims to change? Is the expected change related to protection?
- If not, are there other references in the programme documents to 'do no harm' principles, to protection principles (see Sphere Project, 2012), or to other sectoral and thematic minimum standards for integrating and mainstreaming protection in humanitarian action (e.g. CPWG, 2012; Sutton et al., 2012)?
- Are the drivers of protection needs identified in the assessment, programme documents, or result framework?
- Has the expected beneficiary population of the programme been identified?
- Is the results framework of the programme coherently articulated? Do the outputs, outcomes

31 For a description of monitoring work and how does monitoring can provide different types of information along a result frameworks and logframes, see for instance the IFRC Project/programme monitoring and evaluation guide (IFRC, 2011b).

and goal follow a coherent logic? How does protection feature in the resulting framework (e.g. as a set of specific activities with explicit result? Or is protection integrated in other sectoral interventions?)

- Are the objectives clear and realistic? Are they measurable (quantitatively or qualitatively)? Do they respond to the needs identified?
- Do proposed activities connect to the expected changes and desired results?

EVALUBILITY CHECKLIST 3: Availability of information

During a pre-evaluation process, or during an evaluability assessment exercise, that type of questions that can be asked could be formulated along these lines:

- Has the programme or intervention generated data needed to carry out disaggregated analysis by sex and age (at minimum), and by other characteristics, vulnerabilities, or other lines of affiliation to groups and sub-groups depending on the context and programme evaluated?
- Was the initial programme or intervention design based on disaggregated data, and was this used to develop a protection analysis or other type of assessment and baseline studies?
- Do project/implementing partners (if present) gather and use disaggregated information as part of monitoring day-to-day implementation and mid-course corrections during the life of the project?
- Are there gaps in the data generated by the programme? If yes, is the evaluation expected to generate or reconstruct data to cover for those gaps in order to carry out the analysis and draw evaluative conclusions? Is it realistic to do so with available resources and within the timeframe of the evaluation?

EVALUBILITY CHECKLIST 4: Conduciveness of the context

During a pre-evaluation process, or during an evaluability assessment exercise, that type of questions that can be asked could be formulated along these lines:

- Would the internal conditions of the programme/project and the broader external conditions of the context within which the project is situated allow for an evaluation to take place? Are conditions conducive for ethical, primary data collection and field visits to take place?
- Are there resources, timing and security restrictions that should be taken into account at the scoping and design stage of the evaluation?
- Is there an adequate mix of skills and expertise in the programme ready to 'host' an evaluation mission?
- Are there sufficient human resources available at national/local level for the types of data collection that are to be undertaken? If there are deficiencies, is it possible to invest in developing the skills of the national/local evaluation team members who will undertake these tasks?

Box 16: Template example: Example evaluability checklists template looking at availability of information

This is an example by Cordula Reimann (2012) of a generic template for an evaluability checklist developed in the context of a peace-building initiative.

The checklist looks at the evaluability aspect of availability of information. Such checklist could easily be expanded to include more details around the elements expected to be in place to adjust along the spectrum from lower to higher evaluability.

Expected LOW evaluability conditions	Expected MEDIUM evaluability conditions	Expected HIGH evaluability conditions
Implicit ToC ☐	Implicit but realistic ToC ☐	Explicit ToC ☐
Unclear ToC ☐	Explicit but inappropriate ToC (i.e. ToC contradicts knowledge of peacebuilding practice or is not suited to the context) ☐	Clear and realistic ToC ☐
No baseline ☐	Condensed baseline with data-gathering is focused on a few key indicators for selected goals ☐	Complete baseline ☐
	No baseline but a more comprehensive monitoring at the beginning ☐	
No monitoring system ☐	Monitoring system in place but not used routinely ☐	Monitoring system in place to gather & systematise all necessary information ☐
	Insights from monitoring are not translated into programme changes ☐	
No indicators ☐	Indicators exist, but unrealistic, unmeasurable or unclear ☐	Indicators exists & re SMART ☐
No access to stakeholders & programme participants / programme recipients ☐	Difficult and limited access to stakeholders ☐	Access to stakeholders ☐

The team or the evaluability assessor can tick the respective boxes to indicate where the programme (or other intervention being evaluated) stands in terms of evaluability along the HIGHER-to-LOWER continuum in evaluability status (indicated by the dotted arrow). If most of the ticked boxes are in the left, red column, a programme may not be ready for evaluation in that given moment and it may be useful to: (1) go back to the spectrum of evaluative and reflective options (Figure 3) to consider which type of exercise could be more appropriate; or (2) explore opportunities (outside the remit of evaluation) where the programme can be advised and supported in making changes to those elements identified as weaker in the evaluability analysis.

Source: Adapted from Reimann (2012: 17)

TOOLKIT item #2 – Evaluative rubrics

The key feature that sets evaluation apart from descriptive research from evaluation is that evaluations require us to ask questions about how good something is, and whether it is good enough (Davidson, 2005). Evaluative rubrics are an increasingly common tool used to carry out this type of analysis in evaluation.

Evaluative rubrics are tables that describe what the evidence and indicators should look like at different levels of performance in order to make explicit how judgements are made in an evaluation when assessing the quality, value, or importance of an intervention or programme, policy or service provided. Originally developed and extensively used in the field of education evaluation, rubrics are made up of two main components:

1. The aspects of performance the evaluation focuses on[32]
2. Descriptors that articulate what performance looks like at each level[33](Oakden, 2013: 5).

Why and how can rubrics be helpful to evaluators?

a. They can help evaluators tackle the challenge of 'valuing in evaluation'. This is about answering questions such as: on what basis do we make judgments about performance, quality, and effectiveness? And according to whom? (Julnes, 2012)
b. They can help make transparent how the evaluators apply their professional judgment in order to draw succinct evaluative conclusion and for this reason they have been increasingly discussed in aid evaluation as a conduit to evaluative reasoning[34]. (Davidson, 2005; 2014)
c. They can be used as a 'sense-making' tool because 'as the evidence layers and builds, it is possible to systematically make sense of many streams and lines of evidence, in a concise and cohesive way.' (King et al., 2013: 13)

32 This is also referred to (often confusingly) in evaluation literature as 'evaluative criteria', 'quality distinctions', 'merit criteria', dimensions of merit or indicators.

33 This is also referred to in evaluation literature as 'merit determination' (see Scriven, 1991).

34 King et al. (2013: 20) went as far as arguing: 'We believe rubrics make evaluation accessible and create demand for evaluative thinking well beyond the group of people who think of themselves as evaluators.'

Table 4: Example of ratings used to assess quantitative and qualitative data against each rubric

Rating	Quantitative and qualitative data
Excellent: Always	Clear example of exemplary performance or best practice in this domain; no weaknesses. Likely that 90% or more agree with statement to a considerable or high degree.
Very good: Almost always	Very good to excellent performance on virtually all aspects; scoring overall but not exemplary; no weaknesses of any real consequence. Possibly 80-90% agree with statement to a considerable or high degree.
Good: Mostly, with some exceptions	Reasonably good performance overall; might have a few slight weaknesses but nothing serious. In the range of 60-80% agree with statement to a considerable or high degree, and no more than 15% agree to a limited or very limited degree
Adequate: Sometimes, with quite a few exceptions	Fair performance, some serious but non-fatal weaknesses on a few aspects. Around 40-60% agree with statement to a considerable or high degree, and no more than 15% agree to a limited or very limited degree.
Poor: Never (or occasionally, with clear weaknesses evident)	Clear evidence of unsatisfactory functioning; serious weaknesses across the board on crucial aspects. Probably less than 40% agree with statement to a considerable or high degree. Source: Oakden (2013) originally adapted from Davidson (2005) and reproduced in Davidson (2014: 12)

How is data collected and analysed to populate evaluation rubrics?

There are two broad steps in developing rubrics:

1. The first step is to develop (usually in a participatory manner) the rich descriptions about the different (agreed) performance dimensions (indicators) and make explicit the different levels of performance of the programme or intervention.
2. The second is to consider the different types evidence (qualitative and quantitative) that might be used to draw a conclusion based on the definitions of performance.
3. A well-crafted rubric should paint the picture of what the mix of qualitative and quantitative evidence would look like, and this also gives a clear sense of what will be needed to determine how performance should be rated.

Where existing data is to be used or the evidence has already been gathered, the key is not to define the rubric solely around what is available, but rather to paint the broad picture of what performance looks like regardless of what evidence is available (Davidson, 2014: 6). Examples of how rubrics have been used in evaluation can be found in Oakden (2013).

TOOLKIT ITEM #3 – A partial menu of evaluation approaches and designs

This toolkit item is a partial menu of possible evaluation approaches and designs. It is not intended to be exhaustive but rather to present an initial overview of some of the options that evaluators may consider.

Box 17: Selected descriptive and process-centred approaches to evaluation and specific designs applications and techniques

Partipatory approaches

General features

Participatory evaluation approaches involve stakeholders in all aspects of the evaluation, including technical considerations.

The exercise of power and decision-making within the evaluation process itself shifts from the evaluator to the programme participants themselves. The evaluators' role shifts from expert to facilitator.

Patton (1997) described the basic principles of participatory evaluation as follows:

- Evaluation process involves participants' skills in goal-setting, establishing priorities, selecting questions, analysing data, and making decisions on the data.
- Participants own (commit to) the evaluation, as they make decisions and draw their own conclusions.
- Participants ensure that the evaluation focuses on methods and results they consider important.
- People work together, facilitating and promoting group unity.
- All aspects of the evaluation should be understandable and meaningful to participants.
- Facilitators act as resources for learning; participants act as decision makers and evaluators.

Specific design applications and techniques

Empowerment evaluation (Fetterman, Kaftarian, Wandersman, 1996)

Empowerment evaluation aims to increase the probability of achieving programme success by providing programme stakeholders with tools for assessing the planning, implementation, and self-evaluation of their programme. This is often intended to lead to mainstreaming evaluation as part of the planning and management of the programme/organisation.

Action evaluation

Action evaluation (based on concepts associated with action research) is designed for stakeholders to develop and periodically refine meaningful programme goals and corresponding evaluation criteria throughout the course of their programme. It requires programme stakeholders to explicitly state and periodically revise their collective goals.

Through a series of self-reflections exercises stakeholders determine what they wish to achieve and what success will look like.

Utilisation-focused and developmental evaluation approaches

General features

A variety of methods and approaches to evaluation that focus explicitly on informing decision-making, helping organisations or groups to learn in real time and adapt their strategies to the changing circumstances around them.

Specific design applications and techniques

Patton's Development Evaluation (DE) (Patton, 2011)

- DE is designed and facilitated to provide feedback, generate learning, and either supports strategy decisions or affirms changes to them.
- Choices about whether to use this approach should be based on judgements about the level of independence needed in the evaluation and also the opportunities that exist for engagement between evaluators and the programme over time.
- DE features internal and/or external evaluators who develop long-term relationships with programme participants.
- Evaluators become part of the programme team to ask evaluative questions, bring data and logic to the table, and facilitate evidence-based assessments and decision-making.
- Evaluators who are embedded may be viewed as having less objectivity and neutrality.

Works well with: Complicated and complex strategies that evolve over time, and innovation and pilot initiatives in the test development and testing phase.

Patton's Utilisation-focused evaluation approach (U-FE) (Patton, 2008)

U-FE is a process that can be structured following a 17-step process checklist that starts with assessing and building programme and organisational readiness for UF-E to conclude with follow up with primary intended users to facilitate and enhance use, and meta-evaluation of use.

There is no specific content or method focus, and no specific methods of data collection and analysis.

Instead, U-FE adheres to a set of principles prescribing that the evaluation should be:

- Judged by their utility and actual use
- Situationally responsive
- Negotiated process between evaluators, stakeholders and other evaluation users
- Oriented toward facilitating decision-making about the issues being evaluated
- Facilitated to support the involvement and engagement in the evaluation process and encourage uptake of evaluation findings.

Real-time-evaluation (RTE) (Cosgrave, Ramalingam, Beck, 2009)

The principles underpinning RTEs in humanitarian action (which is where they are mostly commonly used) combine some features of DE and U-FE to ensure responsiveness to the fluid and fast-paced operational environment where humanitarian actors work.

In an RTE, the primary objective is to provide feedback in a participatory way, during fieldwork, to those executing and managing the humanitarian response.

Works well in the context of developing crisis, while response operations are ongoing, and when they are initiated early in an operation.

RTEs require evaluation team members not only evaluate what has been done but also to look at the plausible consequences of what is being done now. RTEs thus have both forward- and backward-looking components.

RTEs' primary stakeholders are the field team and those managing the operation from headquarters. The evaluation team must communicate its findings to the team in the field, few of whom would have time to read a traditional evaluation report.

RTE reports should be finished or nearly finished when the team leaves the field.

Synthesis approaches

General features

One of their strengths is the ability to overcome some weakness of small sample sizes by compiling data from more than one study.

Key requirement: A strict coding protocol ensures consistency in interpretation. Poor coding protocols and coding errors are likely to threaten the validity of the study.

Specific design applications and techniques

Meta-analysis (Labin, 2008)

Meta-analysis is a quantitative tool that combines the results of different studies in order to yield new insights into the nuances surrounding results and changes at both outcome and impact level.

As a statistical method, meta-analysis requires the conversion of qualitative data into quantitative values.

One of meta-analysis strengths is the ability to combine results across studies and samples to produce a better (more accurate, more statistically robust) estimate of the strength and stability of an intervention or of a relationship between two phenomenon of interest.

Summary excerpt from: Corlazzoli and White (2013: 44)

Selected change-centred and theory-based approaches geared toward answering causal questions

Change-centred approaches to evaluation are geared to explore outcome- and impact-level results and changes, and deal with causal inference. They are geared towards answering causal questions and establishing causal inference in evaluation.

The table below gives an overview of evaluation approaches and possible design applications within the realm of change-centred approaches with a view to:

- Specify on which basis the different methods and designs seek to infer causation (with different dominant orientations to establish attribution or contribution).
- Specify which approaches and specific designs and methods applications can work best in evaluation scenarios with small samples ('small n') (following the work by White and Phillips, 2012). 'Small n' evaluation scenarios are likely to be common in evaluating protection in humanitarian action, especially when purposive sampling is used.

Box 18: Selected change-centred approaches and specific designs, applications and techniques

Experimental designs

General features and specific design applications and techniques

Randomised Control Trials are often assumed to provide the strongest design option to control for selection bias in evaluation because the subjects are randomly assigned to the intervention.

They are most suitable for standardised interventions in identical settings with common beneficiaries and limited ranges of intervention modalities (limited variables to assess).

They infer causation based on random assignment of the intervention to groups.

They are less suitable for evaluations with complex causal factors or intentions to contribute to changes in the protection environment. In the 'messy' contexts of EHA protection experimental designs will rarely be feasible or appropriate.

Quasi-experimental designs

General features

- They make no use of randomisation. 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 exist for a given result. Such rival explanation could also be that the two groups were different from the start – and those different features are those that explain what brought about a result (Davidson, 2005: 246).

They are most suitable for:

Standardised interventions in diverse settings, possibly with diverse beneficiaries.

They infer causation based on establishing comparison groups and/or carrying out repeated measurement over time and/or carrying out before and after comparisons. As such, they require a relatively long evaluation timeframe or existence of strong and relevant monitoring data.

Specific design applications and techniques

Regression discontinuity designs (RDD)

- These are a powerful quasi-experimental statistical design where data are compared for a treatment and control group. People are not assigned randomly to groups, but are chosen based on some cut-off value.
- This makes this design suitable for use in humanitarian contexts as the treatment group can be selected by some value.
- Although they are powerful, these designs work best when considerable amounts of data are available. The designs typically compare some dependant variable with the independent variable that is used as the cut off for the treatment.
- For example, if your programme targeted families with a particular household food security score, you could compare the household food security score after six months against the original score, with separate regression lines for above and below the cut-off point.

Before-and-after designs without comparison group

- A before-and-after design gathers data at two time points. The first time point is before the initiation of an intervention. The second is after the intervention has begun. The goal of the design is to examine if the exposure has changed over time and infer whether this is connected to the intervention.
- The point in time when the first measure ('before') and the second measure ('after') are taken varies. There is no standard rule on when this time point should be. It is not uncommon, though, to see time points that are 6 months to a year before and after an intervention.

Non-experimental of 'small n' evaluation approaches – GROUP I

General features

- These involve evaluation approaches and designs that can be used to answer both descriptive and causal questions.
- When used to answer causal questions they tend to focus on explaining the causal mechanisms at work in a given context.
- They infer causation through the use of use narrative and qualitative approaches to build plausible explanation of results.
- Their goal is to explain what has occurred and how it has occurred.
- Approaches below either seek out evidence to substantiate whether a programme's specified theory of change occurred in practice or they do the same for a number of alternative causal hypotheses which outline what might have occurred if causes or assumptions set out in the theory of change had varied.
- They attempt to establish contribution and causation beyond reasonable doubt by collecting evidence to validate, invalidate, or revise the hypothesised explanations, with the goal of documenting the links in the actual causal chain.

Specific design applications and techniques

Theory-Based Approaches to Evaluation (TBE) (Weiss, 2000; Funnell and Rogers, 2011)

- TBE draw from and employ an explicit programme theory that: (1) spells out a set of hypothesised causal linkages between the intervention and desired outcomes provides; and (2) is used as a basis to analyse both attribution and contribution pathways.
- If applied in a critically reflective manner, TBE can help distinguish poor theory from poor implementation (Weiss, 2000).
- While the literature acknowledges that the findings may not 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.
- TBE should not be seen simply as a replacement for experimental and quasi-experimental designs. For high-stakes evaluations with large budgets and extended time lines, the two may be used in conjunction to strengthen causal attributions, provided they are used skilfully. For the everyday evaluator, working under time and budgetary constraints, ideas from both methodologies should be considered in order to build evidence for inferring causality (Cook, 2000 quoted in Davidson, 2000: 25).

Where theories of change are implicit or unarticulated, TBE may benefit from the participation of an external evaluator. Because TBE hinge on the clarity and strength of the theories of change, they are best served by evaluators with knowledge of the subject matter and TBE. TBE are resource intensive and it is most convincing when used in conjunction with other evaluation approaches such as Outcome Identification/measurement and Implementation Evaluation (Rogers, 2012).

No attempt is made to establish intervention and non-intervention groups and causation is inferred on the basis of:

- Identification / confirmation of causal processes or 'chains'.
- Identification and confirmation of supporting factors and causal mechanisms at work in a given context.

Examples of specific design applications in TBE

Contribution analysis (CA)

CA is an analytical tool using the intervention's strategic plan and assessing the contribution story. It is useful when there is no comparison group. It requires a strong theory of change. (See "TOOLKIT item #4 – ODI/RAPID approach to Theories of Change" on page 94 and "TOOLKIT item #6 – Overview of contribution analysis" on page 97 for more step-by-step guidance.)

Realist evaluation (RE) (Pawson and Tilley, 1997)

- According to a realist perspective, programmes can be seen as theories incarnate; when a programme is implemented, it is testing a theory about what actions can help to bring about change (Westhorp et al., 2011).
- Realist Evaluation sets out to establish: (1) an 'inequivocable causal relationship between a programme and outcome(s)'; and (2) that it was, beyond doubt, a programme that caused a given measurable change, and not some other factor(s).
- Programmes are viewed as being akin to open systems in which there are always multiple and competing mechanisms which interact with the surrounding context to produce outcomes. Pawson and Tilley (1997) sum this up as 'mechanisms + context = outcomes'.
- As all mechanisms interact with context, when replicated in new environments programmes cannot be expected to achieve the same outcomes. RE is designed to explain how, and in what circumstances, programmes generate outcomes, by asking 'what works for whom, in what contexts, in what respects and how' (Pawson and Tilley, 1997).

An excellent summary of RE approaches produced by ODI Methods Lab is available in Westhorp (2014).

Case-based approaches

They are most suitable for customised interventions in diverse settings with diverse beneficiaries that use narrative/ qualitative approaches to build plausible explanation of results

Infer causation based on comparisons across and within cases and analytic generalisation based on theory.

Examples of specific design applications in the case-based approaches group

Qualitative Comparative Analysis (QCA) (Rihoux and Ragin, 2009)

QCA is an analytical tool used to compare multiple situations and determining different combinations of causal conditions.

This method is best used when there are multiple case studies with multiple factors to consider and when all factors are known. QCA will usually produce multiple 'causal recipes', relating to the different conjunctions of causal conditions, which produce a given outcome for a certain group of cases.

This technique is most suitable when several scenarios or aspects of an intervention need to be compared or understood. It can work also for 'medium n' evaluations.

Social Network Analysis (SNA) (Emirbayer and Goodwin, 1995)

Social network analysis (SNA) is a methodology used to examine human behaviour and social change by analysing patterns of relations and relationships between individuals, groups, and/or organisations. SNA works to identify individuals or groups that have strong:

- Centrality: those with many relationships
- Prominence: those with the power and ability to influence networks and individuals
- Brokerage: those who can foster entrepreneurial relations or connections between others.

SNA views social relationships in terms of a 'network theory' made up of nodes (representing individual actors or groups within a network with a point) with ties (representing the strength of the relationship or association with a line) (Emirbayer and Goodwin, 1995).

SNA can be used to measure social relationships in crisis, conflict and fragile environments. It can show who is connected to whom and the strength of the relationship within the larger network. It can also help identifying who are the most significant actors or organisations that an intervention should target. It can also show which actors or organisations need support to be able to operate more effectively with others (Corlazzoli and White, 2013).

Non-experimental of 'small n' evaluation approaches – GROUP I

General features

- Group II of small-n approaches have an explicitly participatory orientation when dealing with causal inference.
- Approaches classified in this group are distinguished from the Group I approaches by the fact that they do not set out to address attribution of cause and effect as explicitly as the Group I approaches.
- In general, the Group II approaches place stakeholder participation at the heart of data collection and analysis. They target programme beneficiaries, implementers and other key stakeholders in order to establish what factors are perceived to have been important in producing change; in so doing, they aim to gain an insight into how a programme is performing and the part that it is playing in driving change (White and Phillips, 2012: 13).

Specific design applications and techniques

Most Significant Change (MSC) (Dart and Davies, 2003)

- MSC is a form of participatory monitoring and evaluation that involves collecting stories at the field level and systematically analysing them to identify how stakeholders experience project outcomes and changes in the conflict.
- MSC provides a method for capturing and analysing stories and exploring values behind the preferences for certain changes.
- MSC may provide programme stakeholders and participants with a better understanding of what is and is not being achieved, and even whether the see achievements as valuable and relevant for their situations. Because of its open-ended questions, data can be collected about multiple dynamics or the overall project as a whole, rather than just the intended outcomes (Rogers, 2011).

Success Case Method (SCM) (Brinkerhoff, 2003 and 2008)

- SCM is a particular type of success case study that combines systematic and rigorous case study methodology with storytelling, and reports results that stakeholders can easily understand and believe.
- SCM is a narrative technique based upon naturalistic inquiry and in-depth case-study analysis. It is intended to be a quick and simple evaluation process geared towards understanding whether an initiative (such as a training or educational programme) is actually working.
- SCM sets out to discover whether an intervention is working or not by searching for particularly successful or unsuccessful instances ('success' and 'non-success' cases).
- SCM does not set out to find out about the 'average' participant, but instead intentionally seeks out the very best (and worst) that a programme has produced in order to understand the contribution that the programme has made to results, the role that contextual factors have played in influencing the different outcomes, and the way in which this information can be used to improve programme performance.
- SCM is akin to other methods such as Appreciative Inquiry. (White and Phillips, 2012: 49)

Outcome Mapping (OM) (Earl et al., 2001; Smutylo, 2005; Ambrose and Roduner, 2009)

see "TOOLKIT item #4 – ODI/RAPID approach to Theories of Change" on page 94.

- OM focuses on behavioural change and related outcomes such as capacity development and policy change.
- The focus is on outcomes rather than the achievement of impacts, which are considered too 'downstream' in the results chain.
- Rather than trying to accurately assess any one organisation's contribution to impact OM seeks to look at behaviours, resulting from multiple efforts, in order to help improve the performance of projects, programmes and policies.
- With OM, 'boundary partners' – the individuals, groups and organisations that interact with projects, programmes and policies – are identified. OM assumes that the boundary partners control change more than the intervention itself.
- The focus of OM is people. It represents a shift away from assessing the development impact of a project or programme toward describing changes in the way people behave through actions either individually or within groups or organisations.
- OM provides a way to model what a programme intends to do. It differs from most traditional logic models because it recognises that different boundary partners operate within different logic and responsibility systems. OM can also be used as an end-of-programme assessment tool when the purpose of the evaluation is to study the program as a whole.
- OM proponents believe that many interventions, especially those focusing on capacity development, can better plan for and assess their contributions to development by focusing on behaviour (Morra Imas and Rist, 2009: 196-197).
- OM proceeds through three stages:
    1. Intentional design designates the intended macro-level changes and corresponding strategies.
    2. Outcome and performance monitoring sets a self-assessment framework and data collection tools for the ongoing monitoring of the programme's actions and progress towards results.
    3. Evaluation planning sets the evaluation priorities and develops an evaluation plan.
- OM recognises that multiple, nonlinear events lead to change. OM looks at the logical links between interventions and behavioural change. OM assumes only that a contribution has been made, rather than assuming or attempting to claim attribution. (Rogers, 2011)

Participatory Impact Assessment (PIA) (Catley, Burns, Abede, and Suji, 2008)

- PIA is an extension of Participatory Rural Appraisal (PRA) and involves the adaptation of participatory tools combined with more conventional statistical approaches specifically to measure the impact of humanitarian assistance and development projects on people's lives.
- The PIA approach emphasises the standardisation and repetition of participatory methods, helping to improve the reliability of the information, but ideally leaving enough scope for the open-ended and flexible inquiry typical of PRA.
- Can be used in both small-n and medium-n evaluations and sample sizes.
- Well-designed PIA can assist communities and NGOs to measure impact using their own indicators and their own methods. PIA is designed around eight stages:
    1. Defining the questions to be answered
    2. Defining the boundaries of the project in space and time
    3. Identifying and prioritising locally defined impact indicators
    4. Deciding which method to use and testing it
    5. Deciding which sampling method and sampling size to use
    6. Assessing project attribution
    7. Triangulation
    8. Feedback and verifying results with the community (Catley, Burns, Abede, and Suji, 2008)

Sources: White and Phillips, 2012; Stern, 2008; Bamberger, Rugh, Mabry, 2012; Stufflebeam and Coryn, 2015; Tsui, Hearn and Young, 2014; Mathison, 2005; Rogers, 2012; Morra Imas and Rist, 2009; Chigas, Church, Corlazzoli, 2014.

TOOLKIT item #4 – ODI/RAPID approach to Theories of Change

The **ODI-RAPID approach to developing Theories of Change** follows the principles of Outcome Mapping (OM), which is used for both programming and M&E purposes focusing on behavioural and organisational change. Concretely, this means looking at changes in people's actions and behaviours within their organisations, not at changes in the things that are produced (Shaxson, 2014: 11). The ODI/RAPID approach to OM-infused Theories of Change proceeds through three steps.

First step: Analyse the current context – this includes asking what ideas, interest groups and processes are influencing policy-making.

Second step: Examine for different stakeholders the changes in actions and behaviours that the agency:

a. **Expects to see:** this indicates initial engagement with the intervention – early, positive responses to it.
b. **Would like to see:** this indicates that some initial changes (often called intermediate outcomes) are appearing. At the level of behaviour change, this also indicates that key actors and programme recipients are showing signs that the messages are being taken on board and changing the way things are done.
c. **Would love to see:** this indicates the higher-order changes that the intervention has been aiming towards. At the level of behaviour change, this indicates that the messages have been internalised.

Third step: Identify what the intervention will do, what others will do, and check assumptions about how these are related.

The lines between the different changes are often blurred, and it is often a matter of judgement as to which change falls into which category (Shaxson, 2014: 11; Young, et al., 2014: 27).

Table 5: Example of an Outcome Mapping infused Theory of Change

	General statement of change	Which stakeholders are involved?	Specific indicators
Current context			
Expect to see: early positive response			
Like to see: active engagement			
Love to see: deep transformation in behaviour			

Source: Shaxson (2014: 12). More details on this tool are found in Young et al. (2014) and at www.roma.odi.org/further_resources_on_developing_an_engagement_strategy_to_influence_policy.html.

TOOLKIT item #5 – ODI/RAPID influence and interest matrix

The RAPID team at the Overseas Development Institute (ODI)[35] has developed a simple matrix (Figure 5) to map the stakeholders in policy and research influencing work that can also be used in assessing the ways that different actors in humanitarian protection can be influenced.

The matrix provides a conceptual basis for stakeholder mapping efforts that can help evaluators to systematically assess what drives the interest, influence and actions of different stakeholders or explains their positions in a programme.

A large amount of useful information can be collected and put in a structured form to describe the relationships between different groups of people and how those groups are likely to behave when confronted with the possibility of change (Young et al., 2014: 14).

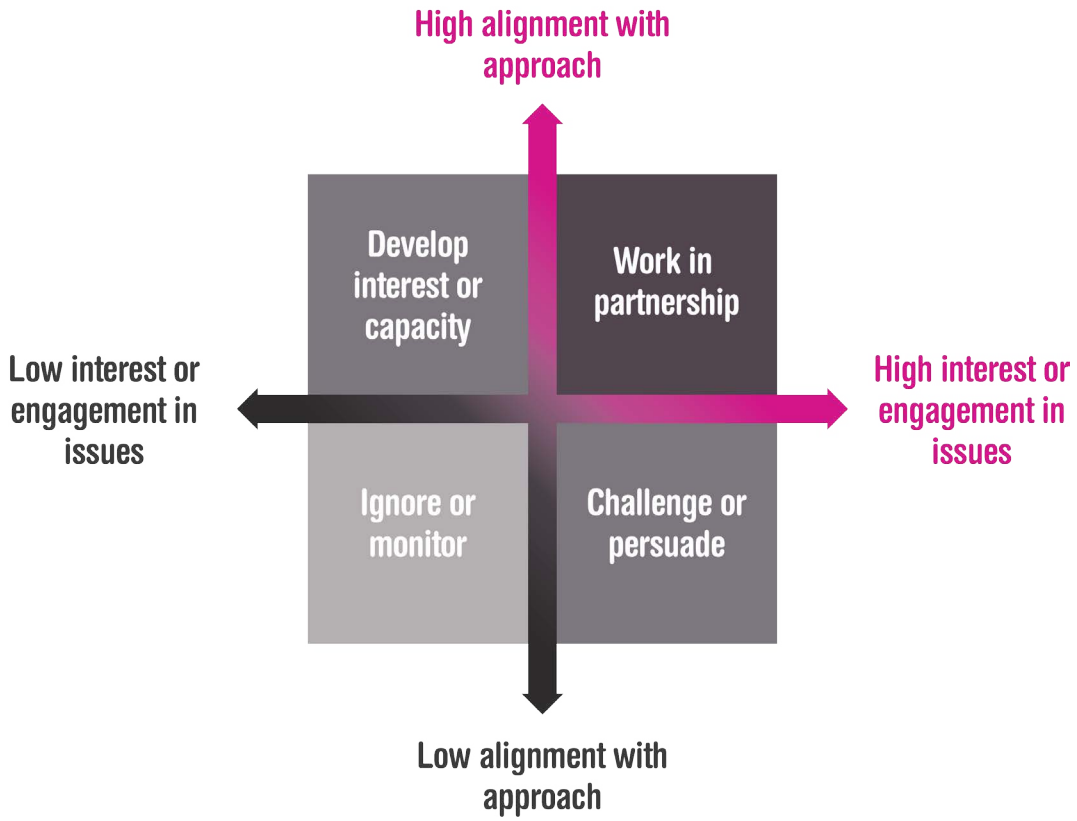
Compared to other stakeholder-mapping tools, this one is noteworthy because not only can help to identify the main stakeholders in an intervention, but also suggest a possible course of action customised towards them (see Mendizabal, 2010 for a step-by-step guide to using the matrix).

In the context of evaluating protection in humanitarian action, the tool can be used in evaluations that look at advocacy and behavioural change components in protection programming. It can also be useful when developing, customising and validating recommendations for different types of actors with varying degrees of interest, alignment and engagement with protection issues in a given context.

The tool can also be used at the analysis and design stage of an intervention. In the original ODI/RAPID formulation, the authors emphasise that 'discussions about who is influential, why and what forms of interest they show in an issue can uncover important relationships between the stakeholders that you can subsequently use to develop your influencing objective. It will also make it more likely that you will consider the full range of people and organisations that need to be included' (Young et al., 2014: 15).

35 See www.odi.org/programmes/rapid

Figure 5: The ODI/RAPID influence and interest matrix



Source: Young et al. (2014): 14

TOOLKIT item #6 – Overview of contribution analysis

Contribution analysis (CA) involves six steps:

1. Develop the theory of change
2. Assess the existing evidence on results
3. Assess alternative explanations
4. Assemble the performance story
5. Seek additional evidence
6. Revise and strengthen the performance story.

The steps are briefly described in Table 6 below. In essence, CA involves using evidence from existing assessment, monitoring and periodic evaluations to see what this data can reveal about the outcomes (or even impacts) of an intervention, while also considering what else besides the intervention could have brought about those results. A provisional performance story is developed from the existing data and should say something about: (1) the extent to which it is reasonable to assume that the programme/project's actions could have contributed to the observed outcomes; and (2) the possible areas of weaknesses and where addition data would be useful.

Developing performance stories can be 'a powerful way of using existing data to determine what is known and where data is needed from additional forms of M&E, or if necessary from an impact evaluation, to provide a more convincing picture' (Perrin, 2013: 13).

Table 6: The six main steps in a contribution analysis process

Steps in contribution analysis	Description
Step 1: Develop the results chain	Develop the programme theory model/programme logic/results chain describing how the programme is supposed to work. Identify the main external factors at play that might account for the outcomes observed. This programme theory should lead to a plausible association between the activities of the programme and the outcomes sought. Some links in the results chain will be fairly well understood or accepted. Others will be less well understood or subject to explanations other than that the programme was the 'cause.' In this way you acknowledge that attribution is indeed a problem.
Step 2: Assess the existing evidence on results	The results chain should provide a good idea of which intended results (outputs, intermediate and end outcomes) could be measured. What evidence (information from performance measures and evaluations) is currently available about the occurrence of these various results? The links in the results chain also need to be assessed. Which are strong (good evidence available, strong logic, or wide acceptance) and which are weak (little evidence available, weak logic, or little agreement among stakeholders)?
Step 3: Assess the alternative explanations	Outcomes by definition are influenced not only by the action of the programme but also by external factors — other programmes, as well as social and economic factors. In addition to assessing the existing evidence on results, there is a need to explicitly consider the extent of influence these external factors might have. Evidence or logical argument might suggest that some have only a small influence and that others may have a more significant influence on the intended results.
Step 4: Assemble the performance story	With this information, you will be able to set out your performance story of why it is reasonable to assume that the actions of the programme have contributed (in some fashion, which you may want to try and characterise) to the observed outcomes. How credible is the story? Do reasonable people agree with the story? Does the pattern of results observed validate the results chain? Where are the main weaknesses in the story? There always will be weaknesses. These point to where additional data or information would be useful. If getting additional evidence is not possible (at least for now), then this is the most you can say about the extent to which the programme has made a difference.
Step 5: Seek out additional evidence	To improve your performance story you will need additional evidence. This could involve information on both the extent of occurrence of specific results in the results chain and the strength of certain links in the chain. A number of strengthening techniques that you might be able to adopt are outlined in this work.
Step 6: Revise and strengthen the performance story	With the new evidence, you should be able to build a more credible story, one that a reasonable person will be more likely to agree with. It will probably not be fool-proof, but will be stronger and more credible.

Source: Mayne (2011) also cited in Perrin (2013)

Bibliography

Bibliography

Ager, A., Ager, W., Stavrou, V. and Boothby, N. (2011) Inter-agency guide to the evaluation of psychosocial programming in emergencies. New York: UNICEF. (www.alnap.org/resource/10214.aspx).

Alexander, J. and Bonino, F. (2015) Addressing causation in humanitarian evaluation: a discussion on designs, approaches and examples. ALNAP Method Note. London: ALNAP. (www.alnap.org/resource/19453).

Alkin, M. (ed.) (2004) Evaluation roots: Tracing theorists' views and influences. Thousand Oaks, CA: SAGE. (www.alnap.org/resource/22276.aspx).

Allaire, A. (2013) Protection interventions for older people in emergencies. London: HelpAge International. (www.alnap.org/resource/19065.aspx).

ALNAP (2015) The State of the Humanitarian System. ALNAP Study. London: ALNAP/ODI. www.sohs.alnap.org

ALNAP. (2016) Evaluation of Humanitarian Action Guide. ALNAP Guide. London: ALNAP. (www.alnap.org/resource/what-we-do/evaluation/eha).

Ambrose, K. and Roduner, D. (2009) A conceptual fusion of the logical framework approach and outcome mapping. London: OMideas. (www.alnap.org/resource/22277.aspx).

Andersen, O. W., Bull, B. and Kennedy-Chouane, M. (eds) (2014) Evaluation methodologies for aid in conflict. Abingdon: Routledge. (www.alnap.org/resource/22278.aspx).

Anderson, M. B. (1999) Do no harm: How aid can support peace or war. London: Lynne Rienner. (www.alnap.org/resource/8805.aspx).

Anderson, M. B. (ed.) (2000) Options for aid in conflict: Lessons from field experience. Cambridge, MA: CDA. (www.alnap.org/resource/22310.aspx).

Anderson, M.B. and Olson, L. (2003) Confronting war: Critical lessons for peace practitioners. Cambridge, MA: CDA. (www.alnap.org/resource/22279.aspx).

Axinn, W. G., Ghimire, D. and Williams, N. E. (2011) Collecting survey data during armed conflict. Ann Arbor, MI: Population Studies Center. (www.alnap.org/resource/22280.aspx).

Bain, A. and Guimond, M. (2014) 'Impacting the lives of survivors: using service-based data in GBV programmes'. Humanitarian Exchange 60. London: HPN/ODI. (www.alnap.org/resource/11016.aspx).

Bamberger M., Rao V. and Woolcock M. (2010) Using mixed methods in monitoring and evaluation: Experiences from international development. Washington, DC: World Bank. (www.alnap.org/resource/22281.aspx).

Bamberger, M. (2012) Introduction to mixed methods in impact evaluation. Washington DC: InterAction (www.alnap.org/resource/8079.aspx).

Bamberger, M., Rugh, J. and Mabry, L. (2012) RealWorld evaluation working under budget, time, data, and political constraints. London: SAGE. (www.alnap.org/resource/8075.aspx).

Barnett, C., Barr, J., Christie, A., Duff, B. and Hext, S. (2011) Governance and conflict indicators report. Hove: ITAD. (www.alnap.org/resource/22282.aspx).

Barnett, C. and Bennett, J. (2014) 'Critical reflections on the South Sudan evaluation of conflict prevention and peacebuilding activities', in Anderson, O. W., Bull, B., and Kennedy-Chouane, M. (eds) Evaluation methodologies for aid in conflict. Abingdon: Routledge. (www.alnap.org/resource/22278.aspx).

Beck, T. (2006) Evaluating humanitarian action using the OECD-DAC criteria: an ALNAP guide for humanitarian agencies. ALNAP Guide. London: ALNAP. (www.alnap.org/resource/5253.aspx).

Befani, B. (2012) 'Models of causality and causal inference', in Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R. and Befani, B. Broadening the range of designs and methods for impact evaluations. London: DFID. (www.alnap.org/resource/8196.aspx).

Bennett, J., Pantuliano, S., Fenton, W., Vaux, A., Barnett, C. and Brusset, E. (2010) Aiding the peace: A multi-donor evaluation of support to conflict prevention and peacebuilding activities in Southern Sudan 2005-2010. Hove: ITAD. (www.alnap.org/resource/8287.aspx).

Berry, K. (2009) Safety with dignity. A field manual for integrating community-based protection across humanitarian programs. Camperdown: ActionAid Australia. (www.alnap.org/resource/22275.aspx).

Berry, K. and Reddy, S. (2010) 'Making space for community-based protection in the humanitarian protection landscape'. Humanitarian Exchange 46. London: HPN/ODI. (www.alnap.org/resource/11051.aspx).

Böhnke, J., Köhler, J. and Zürcher, C. (2010) Assessing the impact of development cooperation in north east Afghanistan 2005-2009. Bonn: BMZ. (www.alnap.org/resource/22283.aspx).

Böhnke, J., Koehler, J., Zürcher, C. (2014) 'Assessing development cooperation in north east Afghanistan with repeated mixed-method surveys', in Anderson, O. W., Bull, B., and Kennedy-Chouane, M. (eds) Evaluation methodologies for aid in conflict. Abingdon: Routledge. (www.alnap.org/resource/22278.aspx).

Bonino, F. (2014) Evaluating protection in humanitarian action: issues and challenges. ALNAP Working Paper. London: ALNAP. (www.alnap.org/resource/19237.aspx).

Boothby, N., Newman, J., Tanabe, M., Prowitt-Smith, L., Ager, A. and Wessells, M. (2006) Assessment and evaluation of psychosocial programming for crisis-affected children: A good practice initiative. New York: UNICEF. (www.alnap.org/resource/19069.aspx).

Borton, J. and Eriksson J. (2004) Lessons from Rwanda – lessons for today. Assessment of the impact and Influence of the joint evaluation of emergency assistance to Rwanda. Copenhagen: DANIDA. (www.alnap.org/resource/3354.aspx).

Borton. J., Buchanan-Smith, M. and Otto, R. (2005) Support to internally displaced persons: Learning from evaluation. Synthesis report on a joint evaluation programme. Stockholm: SIDA. (www.alnap.org/resource/3244.aspx).

Breckon, J. and Hay, J. (2015) Knowing how to protect: Using research evidence to prevent harm to children. London: Alliance for Useful Evidence. (www.alnap.org/resource/22284.aspx).

Brikci, N. and Green, J. (2007) A guide to using qualitative research methodology. London: MSF and London School of Hygiene and Tropical Medicine. (www.alnap.org/resource/13024).

Brinkerhoff, R. O. (2003) The success case method: Find out quickly what's working and what's not. San Francisco: Berrett-Koehler Publishers. (www.alnap.org/resource/22285.aspx).

Broegaard, E., Bull, B., and Kovsted, J., 'Evaluation approaches in situations of conflict and fragility', in Anderson, O. W., Bull, B., and Kennedy-Chouane, M. (eds) Evaluation methodologies for aid in conflict. Abingdon: Routledge. (www.alnap.org/resource/22278.aspx).

Brown, M., and Perschler, V. (2013) Evaluation of UNICEF programmes to protect children in emergencies. New York: UNICEF. (www.alnap.org/resource/12491.aspx).

Brusset, E., Bak, M., Collin, C., Hansen, A., Douma, N., Elakano, J., Otto, R., Perks, R., Ngalimbaya, S., Ngendakuriyo, A., de Valensart, L., Vlassenroot, K. and Voyadzis, C. (2011) Amani labda, peace maybe. Joint evaluation of conflict prevention and peace building in the Democratic Republic of Congo. Brussels: Belgian Ministry of Foreign Affairs. (www.alnap.org/resource/22286.aspx).

Brusset, E. and Hooghe, I. (2014) 'The case of Congo: An evaluation approach focusing on context', in Anderson, O. W., Bull, B., and Kennedy-Chouane, M. (eds) Evaluation methodologies for aid in conflict. Abingdon: Routledge. (www.alnap.org/resource/22278.aspx).

Catley, A., Burns, J., Abede, D. and Suji, O. (2008) Participatory impact assessment: a guide for practitioners. Medford, MD: Feinstein International Center, Tufts University. (www.alnap.org/resource/8094.aspx).

CDA. (2004) Reflecting on peace practice project. Cambridge, MA: CDA. (www.alnap.org/resource/22311.aspx).

CDA. (2006) Do no harm checklist. Cambridge MA: CDA. (www.alnap.org/resource/22312.aspx).

CDA. (2009) Reflecting on peace practice – participant training manual. Cambridge, MA: CDA. (www.alnap.org/resource/22313.aspx).

Chambers, R., Karlan, D., Ravallion, M. and Rogers, P. (eds) (2009) Designing impact evaluations: different perspectives. New Delhi: International Initiative for Impact Evaluation. (www.alnap.org/resource/8474.aspx).

Chapman, J. and Wameyo, A. (2001) Monitoring and evaluating advocacy: a scoping study. London; ActionAid. (www.alnap.org/resource/22314.aspx).

Chapman, N., Duncan, D., Timberman, D. and Abeygunawardana, K. (2009) Evaluation of donor-supported activities in conflict-sensitive development and conflict prevention and peacebuilding in Sri Lanka. Paris: OECD. (www.alnap.org/resource/22315.aspx).

Chigas, D., Church, M. and Corlazzoli, V. (2014) Evaluating impacts of peacebuilding interventions. London: DFID. (www.alnap.org/resource/19072.aspx).

Church, C. and Shouldice, J. (2003) The evaluation of conflict resolution interventions, part II: Emerging practice and theory. Derry: INCORE. (www.alnap.org/resource/22317.aspx).

Church, C. and Rogers, M. M. (2006) Designing for results: Integrating monitoring and evaluation in conflict transformation programs. Washington, DC: Search for Common Ground. (www.alnap.org/resource/22318.aspx).

Coe, J. and Majot, J. (2013) Monitoring, evaluation and learning in NGO advocacy. Findings from comparative policy advocacy MEL review project. London: ODI. (www.alnap.org/resource/19073.aspx).

Coffman, J. (2010) Monitoring and evaluating advocacy: Companion to the advocacy toolkit. New York: UNICEF. (www.alnap.org/resource/22319.aspx).

Cohen, D., Bhandari Karkara, N., Stewart, D., Rees, N. and Coffman, J. (2010) Advocacy toolkit: A guide to influencing decisions that improve children's lives. UNICEF: New York. (www.alnap.org/resource/19137.aspx).

Coldren, J. R. and Bynum, T. (1989) Evaluating juvenile justice programs: A design monograph for state planners. Champaign, IL: Community Research Associates. (www.alnap.org/resource/22649.aspx).

Collier, D. and Elman C. (2008) 'Qualitative and Mmultimethod research: Organizations, publications, and reflections on integration', in Box-Steffensmeier, J. M., Brady, H. E. and Collier, D. (eds) The Oxford handbook of political methodology. New York: Oxford University Press. (www.alnap.org/resource/22321.aspx).

Cook, T. D. (2000) 'The false choice between theory-based evaluation and experimentation'. New Directions for Evaluation, 87: 27-34. (www.alnap.org/resource/22322.aspx).

Cook, T. D., Scriven, M., Coryn, C. L. S. and Evergreen, S. D. H. (2010) 'Contemporary thinking about causation in evaluation: A dialogue with Tom Cook and Michael Scriven'. American Journal of Evaluation, 31(1): 105-117. (www.alnap.org/resource/22323.aspx).

Cordero, M. and Seymour, C. (2011) Indicateurs et outils de mesure - cluster protection, RDC 2011. Oxford: Oxfam. (www.alnap.org/resource/22324.aspx).

Corlazzoli, V. and White, J. (2013) Back to basics - a compilation of best practices in design, monitoring & evaluation in fragile and conflict-affected environments. London: DFID. (www.alnap.org/resource/10196.aspx).

Corlazzoli, V. and White, J. (2013) Measuring the un-measurable: Solutions to measurement challenges in fragile and conflict-affected environments. London: DFID. (www.alnap.org/resource/22309.aspx).

Corlazzoli, V. (2014) ICTs for monitoring and evaluation of peacebuilding programmes. London: DFID. (www.alnap.org/resource/22325.aspx).

Coryn, C. (2013) 'Experimental and quasi-experimental designs for evaluation'. Training Materials. Washington DC: American Evaluation Association. (www.alnap.org/resource/22326.aspx).

Cosgrave, J. and Buchanan-Smith, M. (2012) Evaluation of humanitarian action. ALNAP Guide. London: ALNAP/ODI. (www.alnap.org/eha).

Cosgrave, J., Ramalingam B. and Beck, T. (2009) Real time evaluations of humanitarian action. ALNAP Guide. London: ALNAP. (www.alnap.org/resource/5595.aspx).

CPWG. (2012) Minimum standards for child protection in humanitarian action. Geneva: CPWG. (www.alnap.org/resource/21369.aspx).

CPWG. (2014) Briefing note to ensure child protection mainstreaming. Geneva: CPWG. (www.alnap.org/resource/22328.aspx).

Dart, J., and Davies, R. (2003) 'A dialogical, story-based evaluation tool: The most significant change technique'. American Journal of Evaluation, 24(2): 137–155. (www.alnap.org/resource/22329.aspx).

Davidson, J. (2000) 'Ascertaining causality in theory-based evaluation', in 'Program theory in evaluation: challenges and opportunities' [special issue]. New Directions for Evaluation, 2000(87): 17–26. (www.alnap.org/resource/19075.aspx).

Davidson, J. (2005) Evaluation methodology basics: the nuts and bolts of sound evaluation. London: SAGE. (www.anap.org/resource/19076.aspx).

Davidson, J. (2012) 'Don't expect quantitative evidence to answer a qualitative evaluation question'. Genuine Evaluation. (www.alnap.org/resource/22332).

Davidson, J. (2013) 'Rubric methodology basics'. Presentation at the meeting of the American Evaluation Association, 18 October, Washington, DC. (www.alnap.org/resource/22330.aspx).

Davidson, J. (2013) AEA/BetterEvaluation Part 5: Understand causes of outcomes and impacts. Washington, DC: AEA. (www.alnap.org/resource/19077.aspx).

Davidson, J. (2014). Evaluative Reasoning. Methodological briefs impact evaluation 4. Florence: UNICEF. (www.alnap.org/resource/22331.aspx).

Davies, A. and Ngendakuriyo, A. (2008) Protection and assistance to the Congolese refugees in Burundi. Oslo: NRC. (www.alnap.org/resource/12245.aspx).

Davies, R. (2013) Planning evaluability assessments – a synthesis of the literature with recommendations. London: DFID. (www.alnap.org/resource/8952.aspx).

Davis, R., Marriott, S-J., Gibson, S. and Haegeman, E. (2012) Evaluability assessment for DFID's empowerment and accountability and gender teams. Bristol: The IDL Group. (www.alnap.org/resource/22625.aspx).

De Sas Kropiwnicki, Z. (2012) Protecting children from violence: a synthesis of evaluation findings. New York: UNICEF. (www.alnap.org/resource/10137.aspx).

Denzin, N., and Lincoln, Y. (eds) (2008) Handbook of qualitative research. Thousand Oaks, CA: SAGE. (www.alnap.org/resource/9010.aspx).

Devereux, S., Roelen, K., Béné, C., Chopra, D., Leavy, J. and McGregor, J. A. (2013) Evaluating outside the box: an alternative framework for analysing social protection programmes. Brighton, UK: CPS/IDS. (www.alnap.org/resource/19078.aspx).

DFID. (2011) How to note: Guidance on using the revised logical framework. London: DFID. (www.alnap.org/resource/22333.aspx).

DFID (2012) Tools for measurement, monitoring and evaluation – in depth focus on surveys. London: DFID. (www.alnap.org/resource/22690.aspx).

DFID. (2013) Tools for measurement, monitoring and evaluation – making conflict, crime and violence data usable. London: DFID. (www.alnap.org/resource/10194.aspx).

DuBois, M. and Barnett, K. (2007) 'Protection'. Dialogues 4. London: MSF UK. (www.alnap.org/resource/19079.aspx).

Duggan, C. and Bush, K. (2014) 'The ethical tipping points of evaluators in conflict zones'. American Journal of Evaluation, 35(4): 485-506. (www.alnap.org/resource/22334.aspx).

Earl, S., Carden, F. and Smutylo, T. (2001) Outcome Mapping: Building learning and reflection into development programmes. Ottawa: IDRC. (www.alnap.org/resource/11835.aspx).

Emirbayer, M. and Goodwin J. (1994) 'Network analysis, culture, and the problem of agency'. American Journal of Sociology, 99(6): 1411-1454. (www.alnap.org/resource/22336.aspx).

Enghoff, M., Hansen, B., Umar, A., Gildestad, B., Owen, M. and Obara, A. (2010) In search of protection and livelihood: socio-economic and environmental impacts of Dadaab refugee camps on host communities. Nairobi: Royal Danish Embassy, Government of Kenya and Norwegian Embassy. (www.alnap.org/resource/6334.aspx).

Faúndez, A. and Weinstein, M. (2014) Guide for the evaluation of programmes and projects with a gender, human rights and intercultural perspective. Colonia Polanco: UN Women. (www.alnap.org/resource/22335.aspx).

Fetterman, D. M., Kaftarian, S. J. and Wandersman, A. (eds) (1996) Empowerment evaluation: Knowledge and tools for self-assessment and accountability. Thousand Oaks, CA: SAGE. (www.alnap.org/resource/22337.aspx).

Funnell, S., and P. Rogers (2011) Purposeful program theory: Effective use of theories of change and logic models. San Francisco, CA: Jossey-Bass. (www.alnap.org/resource/8193.aspx).

Gaarder, M. and Annan J. (2014) 'Impact evaluation for peacebuilding: challenging preconceptions', in Anderson, O. W., Bull, B., and Kennedy-Chouane, M. (eds) Evaluation methodologies for aid in conflict. Abingdon: Routledge. (www.alnap.org/resource/22278.aspx).

Gerring, J. (2011) Social science methodology: a unified framework. 2nd ed. Cambridge: Cambridge University Press. (www.alnap.org/resource/9015.aspx).

Giossi Caverzasio, S. (2001) Strengthening protection in war: a search for professional standards. Geneva: ICRC. (www.alnap.org/resource/19168.aspx).

Glouberman, S. and Zimmerman, B. (2002) Complicated and complex systems: What would successful reform of medicare look like? Ottawa: Commission on the Future of Health Care in Canada. (www.alnap.org/resource/8119.aspx).

Goldwyn, R., and Chigas, D. (2013) Monitoring and evaluating conflict sensitivity. Methodological challenges and practical solutions. London: DFID. (www.alnap.org/resource/22338.aspx).

Goodhand, J., Klem, B., and Sørbø, G. M. (2014) 'Battlefields of method – Evaluating Norwegian peace efforts in Sri Lanka', in Anderson, O. W., Bull, B., and Kennedy-Chouane, M. (eds) Evaluation methodologies for aid in conflict. Abingdon: Routledge. (www.alnap.org/resource/22278.aspx).

GPC. (2016) Child protection in emergencies coordination handbook. Geneva: GPC. (www.alnap.org/resource/22047.aspx).

GPC. (n.d.) Brief on protection mainstreaming. Geneva: GPC. (www.alnap.org/resource/22339.aspx).

GPC. (2015) Global protection cluster strategic framework 2016-2019. Geneva: GPC. (www.alnap.org/resource/22340.aspx).

Grävingholt, J. and Leininger, J. (2014) 'Evaluating statebuilding suppor: learning from experience or judging from assumptions?', in Anderson, O. W., Bull, B., and Kennedy-Chouane, M. (eds) Evaluation methodologies for aid in conflict. Abingdon: Routledge. (www.alnap.org/resource/22278.aspx).

Green, J. C., Caracelli, V. J. and Graham, W. F. (1989) 'Toward a conceptual framework for mixed-method evaluation designs'. Educational Evaluation and Policy Analysis, 11(3): 255-274. (www.alnap.org/resource/22308.aspx).

Griñó, L., Levine, C., Porter, S. and Roberts, G. (eds) (2014) Embracing evaluative thinking for better outcomes: Four NGO case studies. Washington D.C.: InterAction. (www.alnap.org/resource/20629.aspx).

Guerrero, S., Woodhead, S. and Hounjet, M. (2013) On the right track: a brief review of monitoring and evaluation in the humanitarian sector. (www.alnap.org/resource/8211.aspx).

Guthrie, K., Louie, J., David, T. and Foster, C.C. (2005) The challenge of assessing policy and advocacy activities: Strategies for a prospective evaluation approach. Los Angeles: The California Endowment. (www.alnap.org/resource/22341.aspx).

Hallam, A. (1998) Evaluating humanitarian assistance programmes in complex emergencies. London: ODI. (www.alnap.org/resource/8210.aspx).

Healy, J. A. (2005) 'What is strategic learning and how do you develop an organizational culture that encourages it?' The Evaluation Exchange, 11(2): 8. (www.alnap.org/resource/22320.aspx).

Heider, C. (2011) 'Conceptual framework for developing evaluation capacities: building on good practice', in Rist, R., Boily, M.-H. and Martin, F. (eds), Influencing change: evaluation and capacity building. Washington DC: World Bank. (www.alnap.org/resource/8422.aspx).

Hendricks, M. and Bamberger, M. (2010) 'The ethical implications of under-funding development evaluations'. American Journal of Evaluation, 31(4): 549–556. (www.alnap.org/resource/22342.aspx).

Hext Consulting. (2011) Programmatic approaches to conflict crime and violence data. Hove: Hext Consulting. (www.alnap.org/resource/22343.aspx).

HIF. (n.d.) After action review (AAR) guidelines. London: HIF (www.alnap.org/resource/22344.aspx).

Holman, A. (2012) Knowledge, attitude and practice surveys in child protection. Stockholm: Save the Children. (www.alnap.org/resource/22345.aspx).

Hummelbrunner, R. and Jones, H. (2013) A guide for planning and strategy development in the face of complexity. London: ODI. (www.alnap.org/resource/22346.aspx).

Humphreys, M. Sanchez de la Sierra, R. and van der Windt, P. (2012) Social and economic impact of Tuungane: Final report on the effects of a community driven reconstruction programme in Eastern Democratic Republic of Congo. New York: Columbia University and IRC. (www.alnap.org/resource/22347.aspx).

IASC. (2011) IASC operational guidelines on the protection of persons in situations of natural disasters. Geneva: IASC. (www.alnap.org/resource/12828.aspx).

IASC. (2002) Growing the sheltering tree: protecting rights through humanitarian action, programmes and practices gathered from the field. Geneva: IASC. (www.alnap.org/resource/9999.aspx).

IASC. (2007). IASC Guidelines on Mental Health and Psychosocial Support in Emergency Settings. Geneva: IASC. (www.alnap.org/resource/19090.aspx).

IASC-IAHE Group. (2014) Inter-agency humanitarian evaluations of large scale system-wide emergencies (IAHEs) guidelines. New York: IASC/IAHE. (www.alnap.org/resource/12850.aspx).

IASC. (2014) 'The centrality of protection: What does it mean in practice?' Webinar presentation delivered on 16 December. New York: IASC. (www.alnap.org/resource/22348.aspx).

ICRC. (2013) Professional standards for protection work carried out by humanitarian and human rights actors in armed conflict and other situations of violence. 2nd Ed. Geneva: ICRC. (www.alnap.org/resource/19095.aspx).

IFRC. (2011a) IFRC framework of evaluation. Geneva: IFRC. (www.alnap.org/resource/7975.aspx).

IFRC (2011b) Project/programme monitoring and evaluation (M&E) guide. Geneva: IFRC. (www.alnap.org/resource/8542.aspx).

IFRC. (2009) Psychosocial interventions: A handbook. Copenhagen: IFRC/Reference Centre for Psychosocial Support. (www.alnap.org/resource/22417.aspx).

IMF. (2003) Data quality assessment framework factsheet. Washington, DC: IMF. (www.alnap.org/resource/22418.aspx).

Innovation Network. (n.d.) Pathfinder: A practical guide to advocacy evaluation. Washington, DC: Innovation Network. (www.alnap.org/resource/22419.aspx).

InterAction. (2012) Project brief: results-based protection. Washington, DC: InterAction. (www.alnap.org/resource/19096.aspx).

InterAction. (2013) NGO and donor consultations on results-based protection. Washington, DC: InterAction. (www.alnap.org/resource/19097.aspx).

InterAction. (2014) Protection mainstreaming learning event: meeting note. Washington, DC: InterAction. (www.alnap.org/resource/11601.aspx).

Jean, I. and Ernstorfer, A. (2014) Alternative approaches to evaluating peacebuilding initiatives. Boston: CDA. (www.alnap.org/resource/22420.aspx).

Jones, H. and Hearn, S. (2009) Outcome Mapping: a realistic alternative for planning, monitoring and evaluation. London: ODI. (www.alnap.org/resource/22421.aspx).

Julnes, G. (2012a). 'Editor's notes'. New Directions for Evaluation, 2012(133): 1-2. (www.alnap.org/resource/22422.aspx).

Julnes, G. (2012b). 'Managing valuation'. New Directions for Evaluation, 2012(133): 3-15. (www.alnap.org/resource/22422.aspx).

Kee, J. E. (1999) 'At what price? Benefit-cost analysis and cost-effectiveness analysis in program evaluation'. The Evaluation Exchange, 5(2/3): 4-5. (www.alnap.org/resource/22423.aspx).

King, J., McKegg, K., Oakden, J. and Wehipeihana, N. (2013) 'Rubrics: a method for surfacing values and improving the credibility of evaluation'. Journal of MultiDisciplinary Evaluation, 9(21): 11-20. (www.alnap.org/resource/22424.aspx).

Körppen, D., Ropers, N. and Giessmann, H. J. (eds) (2011) The non-linearity of peace processes: theory and practice of systemic conflict transformation. Leverkusen: Verlag Barbara Budrich. (www.alnap.org/resource/22676.aspx).

Kushner, S. (2005) 'Qualitative Control – A review of the framework for assessing qualitative evaluation'. Evaluation, 11(1): 111-122. (www.alnap.org/resource/22626.aspx).

Kusters, C., van Vugt, S., Wigboldus, S., Williams, B. and Woodhill, J. (2011) Making evaluations matter: A practical guide for evaluators. Wageningen: Centre for Development Innovation. (www.alnap.org/resource/22425.aspx).

Labin, S. (2008) 'Research synthesis: Towards broad-based evidence', in Smith, N. and Brandon, P. (eds.) Fundamental issues in evaluation. 89-111. New York: Guilford Press. (www.alnap.org/resource/11040.aspx).

Landis, D., Stark, L., Mansourian, H. and Ager, A. (2013) Examining child protection rapid assessment: a structured review of field learning from the Child Protection Rapid Assessment (CPRA) toolkit. New York: CPC Network. (www.alnap.org/resource/19100.aspx).

Landis, D., Williamson, K., Fry, D. and Stark, L. (2013) Measuring violence against children in humanitarian settings: a scoping exercise of methods and tools. New York and London: CPC Network and Save the Children UK. (www.alnap.org/resource/19101.aspx).

Leeuw, F., and Vaessen, J. (2009) Impact evaluations and development: NONIE guidance on impact evaluation. Washington: NONIE. (www.alnap.org/resource/8194.aspx).

Leviton, L. C., Kettel Khan, K., Rog, D., Dawkins, N. and Cotton, D. (2010) 'Evaluability assessment to improve public health policies, programs, and practices'. Annual Review of Public Health 31(1): 213–233. (www.alnap.org/resource/22636.aspx).

LFA. (2013) Advocacy evaluation mini-toolkit: tips and tools for busy organizations. San Francisco: LFA. (www.alnap.org/resource/8949.aspx).

Mahoney, J. and Goertz, G. 2006. 'A tale of two cultures: contrasting quantitative and qualitative research'. Political Analysis, 14(3): 227-249. (www.alnap.org/resource/22637.aspx).

Marcus R. (2014) Child protection violations and poverty in low- and middle-income countries. London: ODI. (www.alnap.org/resource/22638.aspx).

Marcus, R. and Page, E. (2014) Drivers of change in gender norms: An annotated bibliography. London: ODI. (www.alnap.org/resource/22639.aspx).

Mayne, J. (2001) 'Addressing attribution through contribution analysis: Using performance measures sensibly'. The Canadian Journal of Program Evaluation, 16(1): 1-24. (www.alnap.org/resource/22640.aspx).

Mayne, J. (2011) 'Contribution analysis: Advancing cause and effect,' in Forss, K., Marra, M. and Schwartz, R. (eds), Evaluating the complex: Attribution, contribution, and beyond. New Brunswick: Transaction Publishers. (www.alnap.org/resource/22630.aspx).

Meyer, S., Muhorakeye, L. and Stark, L. (2015) Measuring impact through a child protection index – report of pilot study – Kiziba Camp, Rwanda. Geneva: UNHCR. (www.alnap.org/resource/22641.aspx).

Mendizabal, E. (2010) The alignment, interest and influence matrix (AIIM). London: ODI/RAPID. (www.alnap.org/resource/19252.aspx).

Miles, M. and Huberman, M. (1994) Qualitative data analysis: An expanded sourcebook. 2nd ed. Thousand Oaks, CA: SAGE. (www.alnap.org/resource/22642.aspx).

Morras Imas, L. and Rist, R. (2009) The road to results: designing and conducting effective development evaluations. Washington, DC: The World Bank (www.alnap.org/resource/8470.aspx).

Morris, M. (ed.) (2008) Evaluation ethics for best practice: Cases and commentaries. New York: Guilford Press. (www.alnap.org/resource/22643.aspx).

Morris, M. and Clark, B. (2013) 'You want me to do WHAT? Evaluators and the pressure to misrepresent findings'. American Journal of Evaluation, 34: 57–70. (www.alnap.org/resource/22544.aspx).

Moser, C. (2012) 'Mainstreaming women's safety in cities into gender-based policy and programmes'. Gender & Development, 20(3): 435-452. (www.alnap.org/resource/8955.aspx).

Murray, J. and Landry, J. (2013) Placing protection at the centre of humanitarian action - study on protection funding in complex humanitarian emergencies. Geneva: GPC. (www.alnap.org/resource/8733.aspx).

Niland, N., Polastro, R., Donini, A. and Lee, A. (2015) Independent whole of system review of protection in the context of humanitarian action. Oslo: NRC. (www.alnap.org/resource/20330.aspx).

NRC (2015) NRC Evaluation Policy. Oslo: NRC. (www.alnap.org/resource/21069.aspx).

NZQA. (2009) Policy and guidelines for the conduct of external evaluation and review. Wellington, New Zealand: NZQA. (www.alnap.org/resource/22645.aspx).

O'Neil and Goldschmid, P. (2013) Oxfam's GROW campaign mid point external evaluation. Oxford: Oxfam. (www.alnap.org/resource/22646.aspx).

Ober, H. (2012) Guidance for designing, monitoring and evaluating peacebuilding projects: using theories of change. London: CARE International. (www.alnap.org/resource/22647.aspx).

ODI. (2014) Tools for bridging research and policy: The RAPID context, evidence, links framework. London: ODI. (www.alnap.org/resource/22660.aspx).

OECD. (1991) DAC principles for evaluation of development assistance. Paris: OECD. (www.alnap.org/resource/22648.aspx).

OECD. (2012) Evaluating peacebuilding activities in settings of conflict and fragility: Improving learning for results. Paris: OECD. (www.alnap.org/resource/8057.aspx).

Oxfam. (2004) Oxfam GB's final report for an ECHO funded project on promoting protection and informed decision making for East Timorese refugees in NTT Province, Indonesia. Oxfam GB: Oxford. (www.alnap.org/resource/19112.aspx).

Oxfam. (2007). Impact measurement and accountability in emergencies: The good enough guide. Oxford: Oxfam. (www.alnap.org/resource/8406.aspx).

Oxfam. (2011) Global Protection Review. Oxford: Oxfam. (www.alnap.org/resource/19111.aspx).

Patton, M. Q. (1997) Utilisation-focused evaluation: The new century text. Thousand Oaks, CA: SAGE. (www.alnap.org/resource/11131.aspx).

Patton, M. Q. (2006) 'Evaluation for the Way We Work'. The Nonprofit Quarterly, 13(1): 28–33. (www.alnap.org/resource/22650.aspx).

Patton, M.Q. (2008a) 'Advocacy Impact Evaluation'. Journal of MultiDiscplinary Evaluation, 5(9). (www.alnap.org/resource/22651.aspx).

Patton, M. Q. (2008b) Utilization-focused evaluation: The new century text. 4th ed. Thousand Oaks, CA: SAGE. (www.alnap.org/resource/10060.aspx).

Patton, M. Q. (2012) 'Contextual pragmatics of valuing'. New Directions for Evaluations, 2012(133): 97-108. (www.alnap.org/resource/22422.aspx).

Patton, M. Q. (2014a) 'Preface', in Griñó, L. et al. Embracing evaluative thinking for better outcomes. Washington DC: InterAction. (www.alnap.org/resource/20629.aspx).

Patton, M. Q. (2014b) Qualitative research & evaluation methods - Integrating theory and practice. 4th ed. Thousand Oaks, CA: SAGE. (www.alnap.org/resource/22652.aspx).

Patton, M. Q. (2014c) Week 47: Rumination #3: Fools' gold: the widely touted methodological "gold standard" is neither golden nor a standard. [BetterEvaluation.org] 4 December. (www.alnap.org/resource/22653.aspx).

Pawson, R. and Tilley, N. (1997) Realistic evaluation. London: SAGE. (www.alnap.org/resource/22627.aspx).

Peersman, G. (2014) Overview: Data collection and analysis methods in impact evaluation. Florence: UNICEF Office of Research. (www.alnap.org/ressource/22654.aspx).

Peersman, G., Guijt, I. and Pasanen, T. (2015) Evaluability assessment for impact evaluation guide. London and Camberra: ODI Methods Lab. (www.alnap.org/resource/21030.aspx).

Pereznieto, P., Montes, A., Routier, S. and Langston, L. (2014) The costs and economic impact of violence against children. London: ODI. (www.alnap.org/resource/22655.aspx).

Perrin, B. (2012) Linking monitoring and evaluation to impact evaluation. Washington, DC: InterAction. (www.alnap.org/resource/19267.aspx).

Powell, M. A. (2012) Ethical principles, dilemmas and risks in collecting data on violence against children: A review of available literature. New York: UNICEF. (www.alnap.org/resource/22327.aspx).

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).

Psychosocial Working Group. (2003). Psychosocial intervention in complex emergencies: A framework for practice. Edinburgh: Psychosocial Working Group. (www.alnap.org/resource/22656.aspx).

Puri, J., Aladysheva, A., Iversen, V., Ghorpade, Y. and Brück, T. (2014) What methods may be used in impact evaluations of humanitarian assistance? New Delhi: 3ie. (www.alnap.org/resource/19288.aspx).

Raab, M. and Stuppert, W. (2014) Review of evaluation approaches and methods for interventions related to violence against women and girls (VAWG). London: DFID. (www.alnap.org/resource/22657.aspx).

Ramalingam, B. and Jones, H., with Reba, T. and Young, J. (2008) Exploring the science of complexity: Ideas and implications for development and humanitarian efforts. London: ODI. (www.alnap.org/resource/22658.aspx).

Ramsay C. R., Matowe L., Grilli R., Grimshaw J. M. and Thomas, R. (2003) 'Interrupted time series designs in health technology assessment: lessons from two systematic reviews of behavior change strategies'. International Journal of Health Technology Assessment in Health Care, 19(4): 613-623. (www.alnap.org/resource/22659.aspx).

Reichhold, U. and Binder, A. (2014) 'The metrics and ethics of protecting civilians'. Global Responsibility to Protect, 6(2): 206-226. (www.alnap.org/resource/19114.aspx).

Reichhold, U., Binder, A. and Niland, N. (2013) Scoping study: what works in protection and how do we know? Berlin: GPPi. (www.alnap.org/resource/9156.aspx).

Reimann, C. (2012) Evaluability assessments in peacebuilding programming. Cambridge, MA: CDA. (www.alnap.org/resource/22628.aspx).

Reimann, C., Chigas, D. and Woodrow, P. (2012) An alternative to formal evaluation of peacebuilding: program quality assessment. Cambridge, MA: CDA. (www.alnap.org/resource/22316.aspx).

Reisman, J., Gienapp, A. and Stachowiak, S. (2007a) A guide to measuring advocacy and policy. Baltimore: Annie E. Casey Foundation. (www.alnap.org/resource/19115.aspx).

Reisman, J., Gienapp, A., and Stachowiak, S. (2007b) A handbook of data collection tools: companion to 'a guide to measuring advocacy and policy'. (www.alnap.org/resource/22661.aspx).

REPSSI. (2009) Mainstreaming psychosocial care and support through child participation. Johannesburg: REPSSI. (www.alnap.org/resource/22663.aspx).

REPSSI. (2010) Are we making a difference? Participatory evaluation tools For monitoring and measuring the impact of psychosocial support programmes For children aged 6 to 18. 2nd ed. Johannesburg: REPSSI. (www.alnap.org/resource/22664.aspx).

Rihoux B., and Ragin, C. C. (eds) (2009) Configurational comparative methods. Qualitative comparative analysis (QCA) and related techniques. Thousand Oaks, CA: SAGE. (www.alnap.org/resource/22665.aspx).

Rist, R., Boily, M.-H. and Martin, F. (2011) 'Introduction. Evaluation capacity building: a conceptual framework', in Rist, R., Boily, M.-H. and Martin, F. (eds), Influencing Change: Evaluation and Capacity Building. Washington DC: World Bank. (www.alnap.org/resource/8555.aspx).

Robinson, S., Metzler, J. and Ager, A. (2014) A compendium of tools for the assessment of the mental health and psychosocial wellbeing of children in the context of humanitarian emergencies. New York: Columbia University. (www.alnap.org/resource/22666.aspx).

Rog, D. (2005) 'Design, Evaluation', in Mathison, S. (ed.) Encyclopedia of evaluation. London: SAGE. (www.alnap.org/resource/22667.aspx).

Rogers, M. (2012a) Evaluating impact in conflict prevention and peacebuilding programs. Boston: CDA. (www.alnap.org/resource/22668.aspx).

Rogers, M. (2012) Selecting evaluation approaches - options in peacebuilding evaluation. London: CARE International. (www.alnap.org/resource/22669.aspx).

Rogers, P. (2000) 'Causal models in program theory evaluation', in 'Program theory in evaluation: challenges and opportunities' [special issue]. New Directions for Evaluation, 2000(87): 47–55. (www.alnap.org/resource/19116.aspx).

Rogers, P. (2008) 'Using programme theory to evaluate complicated and complex aspects of interventions'. Evaluation, 14(1): 29-48. (www.alnap.org/resource/22185.aspx).

Rogers, P. (2009) 'Matching impact evaluation design to the nature of the intervention and the purpose of the evaluation', in Chambers et al. (eds.) Designing impact evaluations: different perspectives. New Delhi: 3ie. (www.alnap.org/resource/8474.aspx).

Rogers, P. (2011) 'Implications of complicated and complex characteristics for key tasks in evaluation', in Forss, K., Marra, M. and Schwartz, R. (eds), Evaluating the complex: Attribution, contribution, and beyond. New Brunswick: Transaction Publishers. 33-52. (www.alnap.org/resource/22360.aspx).

Rogers, P. (2012) Introduction to impact evaluation. Washington, DC: InterAction. (www.alnap.org/resource/6387.aspx).

Rogers, P. (2014a) Overview of Impact Evaluation. Florence: UNICEF. (www.alnap.org/resource/22671.aspx).

Rogers, P. (2014b) Theory of Change. Florence: UNICEF. (www.alnap.org/resource/22673.aspx).

Rogers, P. (2014c) Overview: Strategies for Causal Attribution. Florence: UNICEF. (www.alnap.org/resource/22674.aspx).

Rogers, P. and Hummelbrunner, R. (2012) 'Methodological challenges in using program theory to evaluate pro-poor and equity-focused programs', in Segone, M. (ed.) Evaluation for equitable development results. New York: UNICEF. (www.alnap.org/resource/22675.aspx).

Rogers, P., Petrosino, A., Huebner, T. and Hacsi, T. (2000) 'Program theory evaluation: practice, promise and problems'. New Directions for Evaluation. 2000(87): 5-13. (www.alnap.org/resource/19117.aspx).

Rogers, P. and Weiss, C. H. (2007) 'Theory-based evaluation: reflections ten years on'. New Directions for Evaluation. 2007(114): 63-67. (www.alnap.org/resource/22670.aspx).

Saferworld. (2014) Community security handbook. London: Saferworld. (www.alnap.org/resource/22678.aspx).

Sassòli, M., Bouvier, A. and Quintin, A. (2011) How does law protect in war? Geneva: ICRC. (www.alnap.org/resource/22680.aspx).

Save the Children. (2012) Child protection menu of outcome indicators. London: Save the Children. (www.alnap.org/resource/22681.aspx).

Save the Children Norway. (2014) A kit of tools for participatory research and evaluation with children, young people and adults. Oslo: Save the Children Norway. (www.alnap.org/resource/22683.aspx).

SC-CPI Global Team. (2013) Quality of child protection services global outcome indicator handbook. Stockholm: SC-CPI. (www.alnap.org/resource/19120.aspx).

Scharbatke-Church, C. (2011a) 'Evaluating peacebuilding: not yet all it could be', In Austin, B., Fischer, M. and Giessmann H. J. (eds) Advancing conflict transformation. The Berghof handbook II. Opladen/Framington Hills: Barbara Budrich Publishers. (www.alnap.org/resource/22684.aspx).

Scharbatke-Church, C. (2011b) The use of reflecting on peace practice in peacebuilding evaluation. Cambridge, MA: CDA. (www.alnap.org/resource/22685.aspx).

Schlangen, R. and Coe, J. (2014). The value iceberg: weighing the benefits of advocacy and campaigning. Better Evaluation. (www.alnap.org/resource/22686.aspx).

Schmidt, R., Scanlon, J. and Bell, J. (1979) Evaluability assessment - making public programs work better. (www.alnap.org/resource/22687.aspx).

Scriven, M. (1991) Evaluation Thesaurus. 4th ed. Newbury Park, CA: Sage. (www.alnap.org/resource/8480.aspx).

Scriven, M. (1995). 'The logic of evaluation and evaluation practice', in Fournier D. M. (ed.) 'Reasoning in evaluation: Inferential links and leaps'. New Directions for Evaluation, 1995(68): 49-70. (www.alnap.org/resource/22688.aspx).

Scriven, M. (2010) 'Rethinking Evaluation Methodology. Journal of MultiDisciplinary Evaluation, 6(13): 1-2. (www.alnap.org/resouce/22631.aspx).

Scriven, M. (2012). 'The logic of valuing', in Julnes, G. (ed.). 'Promoting valuation in the public interest: Informing policies for judging value in evaluation'. New Directions for Evaluation, 2012(133): 17-28. (www.alnap.org/resource/22422.aspx).

Searle, L. and Sutton, K. (2012) 'Standards to incorporate protection into humanitarian response: do they work?' Humanitarian Exchange 46. London: HPN/ODI. 13-16. (www.alnap.org/resource/11051.aspx).

Seddon, J. (2013) Speaking of social impact, words matter. Palo Alto, CA: Skoll World Forum. (www.alnap.org/resource/22689.aspx).

Slim, H. and Bonwick, A. (2005) Protection – an ALNAP guide for humanitarian agencies. ALNAP Guide. London: ALNAP. (www.alnap.org/resource/5263.aspx).

Small Arms Survey and CARE (n.d.) Response to helpdesk request from DFID Somalia Conflict Crime and Violence Results Initiative. London: DFID. (www.alnap.org/resource/22691.aspx).

Smutylo, T. (2005) Outcome mapping: A method for tracking behavioural changes in development programmes. Rome: ILAC. (www.alnap.org/resource/22692.aspx).

Snowden, D. (2010) Safe-to-fail probes. Singapore: Cognitive Edge. (www.alnap.org/resource/22693.aspx).

Snowden, D. and Boone, M. (2007) 'A leader's framework for decision making'. Harvard Business Review 2007(11). (www.alnap.org/resource/22694.aspx).

Soussan, J. (2008) MSF and protection: pending or closed? Paris: MSF. (www.alnap.org/resource/19122.aspx).

Sphere Project. (2011) The Sphere handbook – humanitarian charter and minimum standards in humanitarian response. Geneva: Sphere Project. (www.alnap.org/resource/10673.aspx).

Stacey R. D. (1992). Managing the unknowable. San Francisco, CA: Jossey-Bass. (www.alnap.org/resource/22695.aspx).

Stacey, R. D. (1996) Complexity and creativity in organizations. San Francisco, CA: Berrett-Koehler. (www.alnap.org/resource/22696.aspx).

Stachowiak, S. (2013) Pathways for change: 10 theories to inform advocacy and policy change efforts. Washington, DC: Center for Evaluation Innovation. (www.alnap.org/resource/22697.aspx).

Stave, S. E. (2011) Measuring peacebuilding: challenges, tools, actions. Oslo: NOREF. (www.alnap.org/resource/22698.aspx).

Stern, E. (2008) 'Critical for whom and connected to what?' Evaluation, 14(2): 249-257. Los Angeles: SAGE. (www.alnap.org/resource/22632.aspx).

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R. and Befani, B. Broadening the range of designs and methods for impact evaluations: Report of a study commissioned by the Department for International Development. London: DFID. (www.alnap.org/resource/8196.aspx).

Stufflebeam, D. (2001) 'Evaluation Models'. New Directions for Evaluation 89. San Francisco, CA: Jossey-Bass. (www.alnap.org/resource/22700.aspx).

Stufflebeam, D. and Coryn, C. (2014) Evaluation theory, models, and applications. 2nd ed. San Francisco: Jossey-Bass. (www.alnap.org/resource/22633.aspx).

Sutton, K., Searle, L. and Lubrano, P. (2012a) Minimum standards for protection mainstreaming. Milton Keynes: World Vision UK. (www.alnap.org/resource/19145.aspx).

Sutton, K., Searle, L. and Lubrano, P. (2012b) Minimum inter-agency standards for protection mainstreaming. Milton Keynes: World Vision UK. (www.alnap.org/resource/19146.aspx).

Tarrow, S. (2009) 'Bridging the quantitative-qualitative divide', in Brady, H. and Collier, D. (eds) Rethinking social inquiry: diverse tools, shared standards. Lanham, MD: Rowman and Littlefield. (www.alnap.org/resource/11071.aspx).

Tschirgi, N. (2009) Building peace: practical reflections from the field. Oslo: NOREF. (www.alnap.org/resource/22702.aspx).

Tsui, T., Hearn, S. and Young, J. (2014) Monitoring and evaluation of policy influence and advocacy. London: ODI/RAPID. (www.alnap.org/resource/12943.aspx).

UN WOMEN. (2009) Guidance note on carrying out an evaluability assessment. New York: UN WOMEN. (www.alnap.org/resource/22705.aspx).

UN WOMEN. (2010) A manager's guide to gender equality – and human rights–responsive evaluation. New York and Ottawa: UN WOMEN and Carleton University. (www.alnap.org/resource/22707.aspx).

UNDP. (2013) Discussion paper: Innovations in monitoring & evaluating results. New York: UNDP. (www.alnap.org/resource/22704.aspx).

UNEG. (2005) Norms for evaluation in the UN system. New York: UNEG (www.alnap.org/resource/8215.aspx).

UNEG. (2008) Ethical Guidelines for Evaluation. New York: UNEG. (www.alnap.org/resource/22708.aspx).

UNEG. (2011) Integrating human rights and gender equality in evaluation. New York: United Nations Evaluation Group. (www.alnap.org/resource/8171.aspx).

UNEG (2011) Integrating human rights and gender equality in evaluation - towards UNEG guidance. New York: UNEG. (www.alnap.org/resource/8171.aspx).

UNEG. (2014) Integrating human rights and gender equality in evaluations. New York: UNEG. (www.alnap.org/resource/19283.aspx).

UNFPA. (2012) Quality assessment of UNFPA decentralized country programme evaluations. New York: UNFPA. (www.alnap.org/resource/ 22717.aspx).

UNICEF. (2015) Child protection resource pack - how to plan, monitor and evaluate child protection programmes. New York: UNICEF. (www.alnap.org/resource/22718.aspx).

UNICEF. (2013) Ethical Research Involving Children. [Website]. Florence: UNICEF. (www.alnap.org/resource/22719.aspx).

UNICEF. (2013) Global Evaluation Reports Oversight System (GEROS). New York: UNICEF. (www.alnap.org/resource/7948.aspx).

Weiss, C. H. (1997) 'Theory-based evaluation: past, present and future'. New Directions for Evaluation, 1997(76): 41-55. (www.alnap.org/resource/22720.aspx).

Weiss, C. H. (2000) 'Which links in which theories shall we evaluate?' New Directions for Evaluation, 2000(87): 35-46. (www.alnap.org/resource/22721.aspx).

Westhorp, G. (2014) Realist impact evaluation: an introduction. London: ODI. (www.alnap.org/resource/19141.aspx).

Westhorp, G., Prins, E., Kusters, C., Hultink, M., Guijt, I. and Brouwers, J. (2011) Realist evaluation: an overview. Report from an expert seminar with Dr. Gill Westhorp. Wageningen UR: CDI. (www.alnap.org/resource/22634.aspx).

WFP. (2012) WFP Humanitarian Protection Policy. Rome: WFP. (www.alnap.org/resource/22722.aspx).

White, H. (2009) Some reflections on current debates in impact evaluation. New Delhi: 3ie. (www.alnap.org/resource/8178.aspx)

White, H. and Phillips, D. (2012) Addressing attribution of cause and effects in small n impact evaluation: towards and integrated framework. New Delhi: 3ie. (www.alnap.org/resource/8205.aspx).

Williams, V. (2012) 'EPE week: Valerie Williams on evaluating environmental education programs'. AEA365: A tip-a-day by and for evaluators. Washington, DC: AEA. (www.alnap.org/resource/22723.aspx).

Woodrow, P. and Chigas, D. (2011) 'Connecting the dots: evaluating whether and how programs address conflict systems', in Körppen, D., Ropers, N. and Giessmann, H. J. (eds) The non-linearity of peace processes: theory and practice of systemic conflict transformation. Berlin: Berghof Foundation. (www.alnap.org/resource/22676.aspx)

WHO. (2007) Ethical and safety recommendations for researching, documenting and monitoring sexual violence in emergencies. Geneva: WHO. (www.alnap.org/resource/22635.aspx).

WHO & King's College London. (2011) The humanitarian emergency settings perceived needs scale (HESPER): Manual with Scale. Geneva: WHO. (www.alnap.org/resource/22724.aspx).

Yarbrough, D., Shulha, L., Hopson, R. and Caruthers, F. (2011) The program evaluation standards: A guide for evaluators and evaluation users. 3rd ed. Thousand Oaks, CA: SAGE. (www.alnap.org/resource/22725.aspx).

Yin, R. K. (1984) Case study research: design and methods. 1st ed. Beverly Hills, CA: SAGE. (www.alnap.org/resource/22726.aspx).

Young, H. and Maxwell, D. (2009) Targeting in complex emergencies - Darfur case study. Medford, MA: Feinstein International Center. (www.alnap.org/resource/11346).

Young, Y., Shaxson, L, Jones, H., Hearn, S., Datta, A. and Cassidy, C. (2014) Rapid outcome mapping approach – a guide to policy engagement and influence. London: ODI. (www.alnap.org/resource/12944.aspx).

Related ALNAP publications

Evaluating Humanitarian Action Pilot Guide

E-Learning: introduction to Evaluating Humanitarian Action

Using Humanitarian Evaluations for a Change

www.alnap.org

ALNAP

ALNAP
Overseas Development Institute
203 Blackfriars Road
London SE1 8NJ
United Kingdom
T + 44 (0)20 7922 0388
F + 44 (0)20 7922 0399
E **alnap@alnap.org**